Identification and functional predication of long non-coding rnas in rainbow trout and cattle

Jian Wang

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Recommended Citation
Wang, Jian, "Identification and functional predication of long non-coding rnas in rainbow trout and cattle" (2016). Graduate Theses, Dissertations, and Problem Reports. 6907.
https://researchrepository.wvu.edu/etd/6907

This Dissertation is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Dissertation in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/or on the work itself. This Dissertation has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact research.repository@mail.wvu.edu.
IDENTIFICATION AND FUNCTIONAL PREDICTION OF LONG NON-CODING RNAs IN RAINBOW TROUT AND CATTLE

JIAN WANG

Dissertation submitted to the
Davis College of Agriculture, Natural Resources and Design at
West Virginia University
In partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Genetics & Developmental Biology

Jianbo Yao, Ph.D., Chair
Daniel Panaccione, Ph.D.
Vagner Benedito, Ph.D.
Jennifer Hawkins, Ph.D.
Beth Cleveland, Ph.D.

Division of Animal and Nutritional Sciences
Morgantown, WV
2016

KEYWORDS: next-generation sequencing, lincRNA, lncRNA, transcriptome, rainbow trout, cattle

Copyright 2016 Jian Wang
ABSTRACT

Identification and Functional prediction of Long Non-coding RNAs in Rainbow Trout and Cattle

Jian Wang

The simplest definition of a long noncoding RNA (lncRNA) is an RNA transcript larger than 200 nucleotides that does not encode for a functional protein product. This definition distinguishes lncRNAs from small regulatory RNAs such as microRNAs (miRNAs), short interfering RNAs (siRNAs), Piwi-interacting RNAs (piRNAs) and other short RNAs. They have emerged as a new class of regulatory transcripts in recent years. Recent advances in sequencing technologies have opened a new horizon for the identification and annotation of this class of RNAs in many species. With the increasing evidence supporting important roles of lincRNAs in diverse processes, a systematic catalog of these RNA transcripts and their expression across tissues in rainbow trout is warranted. Here we report the systematic identification and characterization of lincRNAs in 15 major tissue types of rainbow trout. We analyzed the known genomic features of the identified lincRNAs including transcript length, exon number and spatiotemporal expression specificity. We also used weighted gene co-expression network to assign functionalities to the lincRNAs, which revealed that lincRNAs are expressed in a strong tissue-specific manner, and many of them are highly associated with biological processes specific to that tissue.

Reproductive phases in rainbow trout is crucial as the energy expenditure to address the synthesis and release of oocytes is taxing. Skeletal muscle during reproductive phase act as an endogenous source to address the energy demand, compromising muscle quality. Reduced muscle quality in turn results in reduced fillet and egg quality. Reproduction in female fish starts with the release of steroid hormone estrogen initiating synthesis of vitellogenin in liver. Estrogen (E2) is a steroid hormone that negatively affects muscle growth and protein homeostasis in rainbow trout but the mechanisms associated with this response are not fully understood. To better characterize the effects of E2 in muscle, we identified differentially regulated mRNAs and LncRNAs
in juvenile rainbow trout exposed to E2. Here, we performed next-generation RNA sequencing and comprehensive bioinformatics analyses to characterize the transcriptome profiles, including mRNAs and long noncoding RNAs (lncRNAs), in skeletal muscle of normal and E2 treated rainbow trout. A total of 226 lncRNAs and 253 mRNAs were identified. We identified crucial pathways, including several signal transduction pathways, hormone response, oxidative response and protein, carbon and fatty acid metabolism pathways. Subsequently, functional lncRNA-mRNA co-expression network was constructed, which consisted of 681 co-expression relationship between 164 lncRNAs and 201 mRNAs. Moreover, an lncRNA-pathway network was constructed. A total of 65 key lncRNAs were identified, which regulate 20 significantly enriched pathways. Finally, the function of a novel lncRNA (lnc-OM9822) was predicted, which may be activated by estrogen receptor alpha (ER1) and involve in the estrogen-signaling pathway. Overall, our analysis not only effectively provides insights into the mRNA and lncRNA association with the effect of E2 on skeletal muscle in rainbow trout, but also provides further insights into understanding the molecular mechanism of lncRNAs.

We also performed ab initio assembly of more than 80 million RNA-Seq reads from bovine oocytes, and identified 1,535 transcribed lncRNAs from 1,183 loci. In addition, by comparing with previous studies and NONCODE database, we found 115 (7.6%) of our lincRNAs overlap with previously reported bovine lncRNAs in NONCODE database and 565 reported in previous studies. Furthermore, we calculated the tissue specificity score for each oocyte lincRNA. The results indicate that the majority of oocyte lincRNAs (80%) are tissue-specific. Finally, the predicted function of oocyte specific lincRNAs suggested the involvement of bovine oocyte lincRNAs in oogenesis through regulating their neighboring protein-coding genes.
DEDICATION

In dedication to my parents for making me who I am, and my wife and daughter for supporting me all the way!
ACKNOWLEDGEMENTS

Foremost, I would love to express my deepest gratitude to my advisor Dr. Jianbo Yao for his patience, support and enlightening guidance during my Ph. D study. I never would have imagined how fortunate I would be to have such a wonderful mentor.

I am sincerely grateful to my committee members Dr. Beth Cleveland, Dr. Jennifer Hawkins, Dr. Vagner Benedito and Dr. Daniel Panaccione for serving on my graduate committee and for each providing me their own teachings and support throughout my graduate study.

My sincere appreciation and deepest love goes to Dr. Hao Ma and my lab mates Dr. Lei Wang, Dr. Liyuan Fu, Dr. Jacqelyn Hand, Prasanthi Koganti, Xiaowei Zang, Mingxiang Zhang, Wentao Jiang and Shanshan Yang for their unwavering support and dear friendship.

Lastly, and most importantly, I would like to thank my family for the continuous love, understanding and encouragement that have never been changed.
# TABLE OF CONTENTS

ABSTRACT .......................................................................................................................... ii
DEDICATION ....................................................................................................................... iv
ACKNOWLEDGEMENTS .................................................................................................... v
LIST OF FIGURES ............................................................................................................. viii
LIST OF TABLES ................................................................................................................ ix
LIST OF ABBREVIATIONS ............................................................................................... x

REVIEW OF LITERATURE ................................................................................................. 1
INTRODUCTION ................................................................................................................ 2
NEXT GENERATION SEQUENCING .................................................................................. 4
  1. Roche/454 FLX Pyrosequencing .................................................................................... 4
  2. ABI SOLID Sequencing ............................................................................................... 5
  3. Illumina Sequencing ..................................................................................................... 5
CHARACTERIZATION OF LncRNA .................................................................................... 7
  1. Defining long noncoding RNAs ..................................................................................... 7
  2. Functions of long noncoding RNAs ............................................................................. 10
  3. The strategy for studying IncRNA .............................................................................. 12
THE STUDY OF LncRNA IN RAINBOW TROUT AND CATTLE ...................................... 14
REFERENCES .................................................................................................................... 17

CHAPTER 1 ......................................................................................................................... 25
ABSTRACT .......................................................................................................................... 27
INTRODUCTION ................................................................................................................ 28
MATERIALS AND METHODS ............................................................................................. 30
RESULTS AND DISCUSSION ............................................................................................ 34
CONCLUSION ..................................................................................................................... 39
REFERENCES .................................................................................................................... 41
FIGURES AND TABLES ...................................................................................................... 47

CHAPTER 2 ......................................................................................................................... 56
ABSTRACT .......................................................................................................................... 58
INTRODUCTION .................................................................................................................. 59
RESULTS ............................................................................................................................... 62
DISCUSSION .......................................................................................................................... 68
REFERENCES ......................................................................................................................... 78

CHAPTER 3 .............................................................................................................................. 101
ABSTRACT .............................................................................................................................. 103
INTRODUCTION ...................................................................................................................... 104
MATERIALS AND METHODS .............................................................................................. 106
RESULTS ................................................................................................................................. 108
DISCUSSION ............................................................................................................................ 112
CONCLUSION ......................................................................................................................... 114
REFERENCES ........................................................................................................................ 115

Summary .................................................................................................................................. 125
APPENDIX ............................................................................................................................... 129
LIST OF FIGURES

REVIEW OF LITERATURE - Figure 1. Genomic contexts of lncRNAs................................. 9

CHAPTER 1 - Figure 1. Pipeline used to identify novel lincRNAs........................................... 46
CHAPTER 1 - Figure 2. Structural characteristics of lincRNAs in comparison to protein-
coding transcripts................................................................................................................. 47
CHAPTER 1 - Figure 3. Comparison of the expression levels of lincRNAs and protein-
coding genes........................................................................................................................ 48
CHAPTER 1 - Figure 4. Tissue-wise distribution of predicted novel lincRNAs.................... 49
CHAPTER 1 - Figure 5. Tissue specificity of lincRNAs and protein-coding genes.............. 50
CHAPTER 1 - Figure 6. Validation of expression specificity of lincRNAs by RT-PCR
analysis................................................................................................................................... 51
CHAPTER 1 - Figure 7. Correlation of expression patterns between pairs of neighboring
genes....................................................................................................................................... 52
CHAPTER 1 - Figure 8. Function prediction of rainbow trout lincRNAs......................... 53

CHAPTER 2 - Figure 1. The number of differentially expressed genes in rainbow trout
under effect of E2......................................................................................................................... 92
CHAPTER 2 - Figure 2. Differentially expressed lncRNAs and mRNAs validated by
qRT-PCR.................................................................................................................................... 93
CHAPTER 2 - Figure 3. The GO analysis of differentially expressed mRNAs.................... 94
CHAPTER 2 - Figure 4. KEGG pathway analysis of differentially expressed genes......... 95
CHAPTER 2 - Figure 5. lncRNA-mRNA co-expression network......................................... 96
CHAPTER 2 - Figure 6. lncRNA-pathway network................................................................. 97
CHAPTER 2 - Figure 7. The key lncRNAs and corresponding crucial pathway................ 98
CHAPTER 2 - Figure 8. Putative motifs across sequence of Inc-OM9822......................... 99

CHAPTER 3 - Figure 1. Total number of putative oocyte lincRNA per chromosome. 120
CHAPTER 3 - Figure 2. Venn diagram of comparative analysis of lincRNA in present
study with bovine lncRNA from similar studies................................................................. 121
CHAPTER 3 - Figure 3. Tissue-specific expression of oocyte lincRNAs......................... 122
CHAPTER 3 - Figure 4. The GO analysis of neighboring mRNAs................................. 123
LIST OF TABLES

REVIEW OF LITERATURE - Table 1. Comparison of different sequencing platforms... 7

CHAPTER 1 - Table 1. Summary of samples and RNA-Seq data. .......................... 54

CHAPTER 2 - Table 1. Summary of samples and RNA-Seq data. ......................... 88
CHAPTER 2 - Table 2. Statistics of transcriptome assembly................................ 89
CHAPTER 2 - Table 3. Most enriched representative GO terms of up- or down-regulated genes response to E2 ........................................................................... 90
CHAPTER 2 - Table 4. Top 20 nodes with the highest degree in differential lncRNA-mRNA co-expression network ................................................................. 91

CHAPTER 3 - Table 1. Statistical summary of bocine oocyte sequencing, assembly and annotation .................................................................................................. 119
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>lincRNA</td>
<td>Large intergenic non-coding RNA</td>
</tr>
<tr>
<td>IncRNA</td>
<td>Long non-coding RNA</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>CDS</td>
<td>Coding DNA sequence</td>
</tr>
<tr>
<td>XIST</td>
<td>X-inactive specific transcript</td>
</tr>
<tr>
<td>PPC2</td>
<td>Phosphoenolpyruvate carboxylase2</td>
</tr>
<tr>
<td>RNA-Seq</td>
<td>RNA sequencing</td>
</tr>
<tr>
<td>FPKM</td>
<td>Fragments Per Kilobase of transcript per Million mapped reads</td>
</tr>
<tr>
<td>BP</td>
<td>Biological process</td>
</tr>
<tr>
<td>CC</td>
<td>Cellular component</td>
</tr>
<tr>
<td>MF</td>
<td>Molecular function</td>
</tr>
<tr>
<td>GV</td>
<td>Germinal vesicle</td>
</tr>
<tr>
<td>MII</td>
<td>Metaphase II</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>nt</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>JS</td>
<td>Jensen-Shannon</td>
</tr>
<tr>
<td>PCC</td>
<td>Pearson correlation coefficient</td>
</tr>
<tr>
<td>PP</td>
<td>Phosphatase</td>
</tr>
<tr>
<td>CDK1</td>
<td>Cyclin dependent kinase-1</td>
</tr>
<tr>
<td>ROSC</td>
<td>Roscovitine</td>
</tr>
<tr>
<td>PCM</td>
<td>Pericentriolar material</td>
</tr>
<tr>
<td>ECM</td>
<td>Extracellular matrix</td>
</tr>
<tr>
<td>FKA</td>
<td>Focal adhesion kinase</td>
</tr>
<tr>
<td>FAs</td>
<td>Focal adhesions</td>
</tr>
<tr>
<td>ERE</td>
<td>estrogen response element</td>
</tr>
<tr>
<td>ER1</td>
<td>estrogen receptor alpha</td>
</tr>
<tr>
<td>CASQ1</td>
<td>Calsequestrin-1</td>
</tr>
</tbody>
</table>
REVIEW OF LITERATURE
INTRODUCTION

The recent advent of next generation RNA sequencing (RNA-Seq) and publication of reference genome for many organisms have allowed researchers to study transcription profile genome wide. It is recently revealed that the vast majority of the mammalian genome (up to 80%) is transcribed, while only 2-3% of the mammalian genome is transcribed into protein-coding RNAs (mRNAs) \(^1\, ^2\). While some of ncRNAs are processed to generate small RNAs including microRNA (miRNA), small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs) and piwi-interacting RNAs (piRNAs), most transcribed ncRNAs are larger than 200 nucleotides in their mature forms and defined as long non-coding RNAs \(^3\, ^5\). More and more lncRNAs are reported to play critical roles in various biological processes, including chromatin modification, regulation of transcription, influence of nuclear architecture and regulation of gene expression on post-transcriptional and post-translational levels \(^5\, ^{14}\). Dysfunction of lncRNAs can lead to a variety of human diseases including cancer \(^{15}\, ^{16}\).

Rainbow trout (Oncorhynchus mykiss) is common fresh water fish cultivated in America, with its importance in supply of aquatic food in the USA and worldwide. In addition, rainbow trout is being used as a model in research in fields like carcinogenesis, immunology, toxicology and evolutionary biology \(^{17}\). To improve aquaculture production and facilitate scientific study, a large amount of study and genetic information has been accumulated which include linkage map \(^{18}\, ^{19}\), physical map \(^{20}\) and genome reference \(^{21}\). However, complete understanding of the biological processes of rainbow trout is far from archived. Identifying lncRNA in rainbow trout would contribute to the current repertoire of rainbow trout lncRNA and genome annotation, and help further improve our understanding of the mechanism of biological processes and the evolutionary importance of these transcripts. Additionally, recent study revealed that most of the important genetic markers, like SNPs, located out of the protein-coding region where mutations could contribute to important economic traits \(^{7}\).

Cattle (Bos taurus) is one of most commonly raised livestock for meat, milk and other dairy products. There are a number of studies reporting bovine lncRNAs across many tissues \(^{22}\, ^{25}\). However, very little is known about the identity and characteristics of
IncRNAs in bovine oocytes. The developmental competence of an oocyte, also known as egg quality, is defined as the ability of the egg to be fertilized and subsequently develop into a normal embryo. Mammalian oocytes harbor a vast collection of RNA and proteins that regulate subsequent early embryonic development and activation of embryonic genome activation. Several studies have reported that IncRNAs play critical roles in the embryonic stem cell regulatory\textsuperscript{3,26-28}. For instance, more than 100 lincRNA promoters were identified to be bound by stem cell factors such as OCT4 and Nanog\textsuperscript{28}. Therefore, the study of IncRNA in bovine oocyte could help understand the embryonic development.

This chapter focuses first on review of the next generation sequencing (NGS). The second part of this chapter is a review of IncRNA characterization, functions and the strategy for studying IncRNA. The last part will review the current study status of IncRNA in rainbow trout and cattle and the necessity of our present study on IncRNA.
NEXT GENERATION SEQUENCING

Chain termination and fragmentation techniques were developed by Sanger in 1970s\textsuperscript{29}, which is called Sanger sequencing and considered as a first-generation sequencing technology. However, the limitation of the Sanger sequencing is the low throughput (up to 96 reads/run). A growing demand of high throughput led to the development of next-generation sequencing (NGS) technologies. Compared with the Sanger sequencing, the new sequencing methods need NGS library to be prepared instead of cloning DNA fragments into bacteria. Second, millions of sequencing reaction are produced parallel instead of a hundred for Sanger sequencing. Furthermore, not like Sanger sequencing, electrophoresis is unnecessary for NGS. Among those powerful and evolving technologies, three platforms are mainly used and widespread for massively parallel DNA sequencing at present: 454 Gnome Sequencer released in 2005 by 454 Life Sciences (now Roche)\textsuperscript{30}, the Solexa/Illumina sequencing platform (Illumina acquired Solexa in 2007) and the ABI SOLID (now Life Technologies). The read length, maximum throughput, cost and runtime varies among the three platforms.

1. Roche/454 FLX Pyrosequencing

Roche/454 pyrosequencing sequencing was the first next-generation sequencing platform to achieve commercial introduction\textsuperscript{30}. The principle of this approach is based on “sequencing by synthesis”. The 454 pyrosequencing relies on the detection of pyrophosphate release along with nucleotide incorporation instead of chain termination with dideoxynucleotides which is used by Sanger sequencing. In each sequencing cycle, only one out of the four possible nucleotides (A/T/G/C) are added so that only one letter can be incorporated on the sequencing template along with release of a pyrophosphate (ppi). The chemical signal (ppi) can be converted to light signal after a series of enzyme reactions. Unincorporated nucleotides are degraded and the next cycle starts with another nucleotide. The first step of sequencing is the library construction, in which each bead carrying oligonucleotides complementary to the 454-specific adapter sequences is associated with a single fragment. Each of fragment:bead complex is embed into individual oil: water micelles that also contain PCR reactants. Then around one million
copies of each DNA fragment are produced on the surface of each bead after PCR amplification. The emulsion shell is then broken and the clonally amplified single molecule beads are ready for loading onto the fibre-optic chip for sequencing. The chip is loaded with one bead per well and the enzymes that catalyze the downstream pyrosequencing reaction steps. The CCD camera that records the light emitted at each bead spot during each sequencing circle.

2. ABI SOLiD Sequencing

The SOLiD platform is another next-generation sequencing platform, which performs the sequencing by ligation. This platform uses an emulsion PCR approach with small magnetic bead to amplify the fragments for sequencing. The probes are designed with known first two nucleotides followed by specific fluorescent dye. A sequencing primer is hybridized to the adapter and its 5’ end is available for ligation to an oligonucleotide hybridizing to the adjacent sequence. Each ligation step is followed by fluorescence detection. Fluorescent dye is excited only if the probe anneals with the template, and then the next ligation cycle starts. After the full amplification of the template, primer should be reset with primer with 1 nucleotide shorter than previous one, and ligation cycles should be repeated again. After several repeats of primer reset, DNA sequence can be extracted by decoding the fluorescent color. SOLiD applied the two-base encoding for base-calling and read each base two times which enables this platform more widely used in genome re-sequencing and SNP discovery due to its highest accuracy.

Briefly, the amplification products are transferred onto a glass surface where sequencing occurs by sequential rounds of hybridization and ligation with 16 dinucleotide combinations labeled by four different fluorescent dyes (each dye used to label four dinucleotides).

3. Illumina Sequencing

The Illumina platform is the most successful and widely-adopted next-generation sequencing platform worldwide. Currently, the main system in Illumina is HiSeq. The most important technique of Illumina platform is the use of labeled nucleotides with reversible terminators to sequence a single base at a time. Illumina sequencing
technology works in three basic steps: amplification, sequencing and image analysis. The nucleic acid samples are sheared into small pieces with adaptors ligated at both ends. The DNA is then loaded onto a specialized chip where bridge amplification and sequencing will take place. Bridge amplification create hundreds of identical DNA called cluster that make sure the fluorescent signal strong enough for detection. During sequencing cycle, all four labeled reversible terminators, primers and DNA polymerase enzyme are added. After the first cycle of amplification, elongation is stopped because of the blocked 3’ terminus of newly added nucleotide. Laser system capture the image of emitted fluorescence from each cluster on flow cell so that the first base for each cluster is recorded. The next cycle can be started following the remove of 3’ terminus. In addition, a semiconductor sequencing technology Ion Torrent has emerged as a new choice for sequencing. This platform is very similar to 454 sequencing except that proton level is measured during nucleotide incorporation instead of pyrophosphate. Another advantage is that no imaging technology is required.

The third generation sequencing technology, such as Pacific bioscience (PacBio) and Nanopore, have great potential to be widely applied in sequencing field with its ability to produce extremely long read. The most attractive feature of PacBio is the use of SMRT (Single molecule real time). PCR amplification bias can be avoided, as the sequencing process doesn’t require PCR procedure. Moreover, the signal produced by this platform is monitored as the same time as the incorporation of nucleotide into the complementary strand\textsuperscript{31}. Taken together, the advent of new sequencing technologies allows the high-throughput data produced with rapidly dropping cost. Here we list the comparison of the different sequencing platforms.
Table 1. Comparison of different sequencing platforms

<table>
<thead>
<tr>
<th>Platform</th>
<th>Roche 454</th>
<th>Illumina</th>
<th>ABI SOLID</th>
<th>Ion Torrent</th>
<th>Pac Bio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principle</td>
<td>Sequencing by Synthesis</td>
<td>Sequencing by Synthesis</td>
<td>Sequencing by Ligation</td>
<td>Single molecule sequencing</td>
<td>Semiconductor Sequencing</td>
</tr>
<tr>
<td>Sequencing chemistry</td>
<td>Pyrosequencing</td>
<td>Reversible terminator</td>
<td>Ligation based</td>
<td>Real-time</td>
<td>Chemical to digital</td>
</tr>
<tr>
<td>Read length (bp)</td>
<td>1000</td>
<td>300</td>
<td>75</td>
<td>400</td>
<td>20000</td>
</tr>
<tr>
<td>Throughput per run (Gb)</td>
<td>0.7</td>
<td>1800</td>
<td>320</td>
<td>10</td>
<td>0.5</td>
</tr>
<tr>
<td>Run time (Hours)</td>
<td>10</td>
<td>27</td>
<td>336</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Cost per 1 million bases (in US$)</td>
<td>10</td>
<td>0.05-0.15</td>
<td>0.13</td>
<td>1</td>
<td>0.75-1.50</td>
</tr>
<tr>
<td>Pros</td>
<td>Long reads</td>
<td>Most widely used</td>
<td>Highest accuracy</td>
<td>Fast run time</td>
<td>Longest read length</td>
</tr>
<tr>
<td>Cons</td>
<td>Low throughput</td>
<td>Short read length</td>
<td>Long run time</td>
<td>Lower throughput</td>
<td>High error rate</td>
</tr>
</tbody>
</table>

CHARACTERIZATION OF LNCRNA

1. Defining long noncoding RNAs

The simplest definition of a lncRNA is an RNA transcript larger than 200
nucleotides that does not encode for a functional protein product. Like the protein-coding mRNAs, IncRNAs are transcribed by RNA Polymerase II, and therefore have features such as 5’-cap, poly-adenylation and alternative splicing\(^3\). LncRNA also displays histone modification signal at promoter and transcribed region, which often are indicative of their expression status\(^3\). According to the genomic position of the locus from which they are transcribed and their proximity to protein coding genes in the genome, IncRNAs can be divided into five categories: sense, antisense, bidirectional, intronic and intergenic IncRNAs (Fig. 1). Known examples of intergenic IncRNA include Xist\(^33\), H19\(^34\), HOTAIR\(^35\) and MALAT1\(^36\). There are also a number of well-documented antisense IncRNAs, such as Tsix\(^37\), Kcnq1ot1\(^38\) and Air\(^39\). COLDAIR is an intronic IncRNA, which is located in the first intron of the flowering repressor locus FLC\(^40\).

Recent studies revealed that the vast majority of the mammalian genome (up to 80%) is transcribed, while only 2-3% of the mammalian genome is transcribed into protein-coding RNAs (mRNAs)\(^1\). It is revealed that most of the transcribed non-coding genomic region encodes for IncRNA. So far, three major IncRNA database including LNCipedia\(^41\), GENCODE and NONCODE\(^42\) have identified more than 100,000 human IncRNA genes. While the abundance of IncRNAs identified, this number represents a conservative lower estimate, since many single exon transcripts and non-polyadenylated transcripts are omitted from this category.
The function of lncRNAs have been a topic of debate since lncRNA sequences lack strong conservation. It is known that mRNAs have selection pressure to conserve the codon usage in open reading frame (ORF) to prevent frameshift mutations and keep the normal function. However, lncRNA sequences showing generally low conservation between species may support lncRNAs as transcriptional noise. On the other hand, an explanation for the low sequence conservation of lncRNAs is that selection may conserve only short region of lncRNAs that are important for structure or sequence-specific interactions. An example showing this conservation flexibility is the lncRNA Xist that silences one of the two X chromosomes in all eutherian females. Xist has several repetitive regions in its sequence and shows very little sequence conservation, indicating the fact that high degree of sequence conservation is not essential requirement for lncRNA functionality. Moreover, in contrast to the low sequence conservation in
lncRNA sequence, the promoters of lncRNAs were reported to be very high sequence conserved indicating that the level of the transcription of lncRNA is highly conserved, although the lncRNA sequences themselves are not\textsuperscript{45}.

2. Functions of long noncoding RNAs

Compared to the large amount of lncRNA identified, only less than 200 lncRNAs were assigned with explicit function for human in lncRNAdb\textsuperscript{46}. Nevertheless, those well-studied lncRNAs implicate their diverse functions in various biological processes. The summarization of our latest understanding of lncRNAs will help us for the further study of lncRNA.

Dosage compensation and genomic imprinting

The first and best-characterized examples of lncRNAs are those involved in dosage compensation and genomic imprinting. Both two processes rely on the formation of silence chromatin. It is known that the difference of gene dosage on X chromosome between XX females and XY males in therian mammals in compensated through the inactivation of one of X chromosome in females\textsuperscript{47}. In placental mammals, X chromosome inactivation is mainly regulated by a 17kb long X (inactive)-specific transcript (Xist) which is transcribed from a cluster of lncRNA loci known as X-inactivation center\textsuperscript{48}. Moreover, lncRNAs are also important in genomic imprinting, the process by which a gene is transcribed from one of its parent allele, either paternal or maternal allele\textsuperscript{49}. Imprinted genes are generally clustered together, which normally contain protein-coding genes and lncRNAs that are reciprocally expressed, such as Kcnq1ot1\textsuperscript{38} and Air\textsuperscript{50} overlapping with Kcnq1ot1 and Igf2r, respectively. Some of these lncRNAs control the imprinted expression of neighboring genes by recruiting chromatin-remodeling factors, such as PRC2 and G9a\textsuperscript{38,51}.

Roles in development

Embryonic development begins with the formation of a group of pluripotent stem cells, which give rise to all cell types of the body. This process needs to be well-regulated. LncRNAs are thought to play important roles in regulating gene transcription. A few lncRNAs were documented as main regulators during development. For example,
Dlx6os1 (the coactivator lncRNA Dlx6 opposite strand transcript 1) is an lncRNA transcribed from antisense of Dlx6 gene which is a transcription factor that have a role in both forebrain and craniofacial development. Dlx6os1 negatively regulates the expression of Dlx6 gene in trans through recruiting methyl-CpG-binding protein 2 (MECP2, which is a repressor) to the gene target. In neurogenesis, the expression of lncRNA-N1 and lncRNA-N3 were shown to recruit PRC2 to REST gene (RE1-silencing transcription factor) which act as repressor of neural genes in non-neuronal cells, thereby promoting neurogenesis. In addition to crucial roles in brain development, lncRNAs are known to function in the development of diverse organs. Fendrr is a lncRNA that regulates heart development. The knockout of Fendrr resulted in embryonic fatality due to impaired heart function. In addition, many lncRNAs have been shown to play key a role in skeletal muscle differentiation. LINCMD1 was reported to control muscle differentiation of mouse myoblasts in vitro through competitively decoying miR-133 and miR-135, thus enhancing the stability of transcription factor of muscle genes.

Hox genes, also known as homeotic genes, are a group of genes that control the body plan along the head-tail axis during embryonic development. In human, 39 HOX genes are grouped into four clusters. HOTTIP is an lncRNA transcribed from 5’ end of the human HOXA locus upstream of HOXA13. The mechanism by which HOTTIP regulates HOXA gene expression relies on its interaction with MLL1 complex (activating complex) and on the formation of chromatin loops between HOTTIP transcripts and various HOXA gene promoters. HOXAIR is another lncRNA transcribed from HOX cluster. The expression of HOXAIR from HOXC cluster recruit PRC2 complex to HOXD cluster and suppress HOXD gene expression in trans.

Pluripotency associated lncRNAs

Several lncRNAs associated with pluripotency have been identified. Notably, these lncRNAs show correlated expression with pluripotency protein coding genes, such as OCT4, NANOG and SOX2. For example, both lncRNA-ES2 and lncRNA-ES1 were reported to interact with repression complex and SOX2, which indicate that these two lncRNAs function as scaffolds to recruit repression complex to the target of SOX2 gene.
Diseases associated lncRNAs

Several lncRNAs are associated with diseases, most notably cancer\textsuperscript{57}. A recent study discovered a lncRNA called Pcat-1 that promotes cell proliferation through interacting with PRC2 in prostate cancer\textsuperscript{58}. MALAT1 is another lncRNA associated with various cancers\textsuperscript{36}. Some lncRNAs are involved in regulation of cell circle through p53 and RB pathways. LncRNA-p21 and PANDA are two lncRNAs found to function in p53 pathway. The former inhibits the expression of target genes involved in apoptosis\textsuperscript{59}. The latter interacts with transcription factor NF-YA to limit expression of proapoptotic genes\textsuperscript{60}.

3. The strategy for studying lncRNA

Identification of lncRNA

The first step of lncRNA study is to identify lncRNAs expressed in the organism of interest. RNA-seq and microarray are two mainly techniques used to identify lncRNAs to date. RNA-seq deep sequencing is a more direct method to reconstruct an entire transcriptome and estimate the abundance of all transcripts. To distinguish lncRNAs locus from protein coding genes, various pipelines were used in previous studies. Computationally, algorithms that calculate coding potential, such as CPC, CPAT and PhyloCSF were used to differentiate coding RNAs from lncRNAs. CPC and CPAT are based on the length and quality of ORF (open reading frame) to assess a transcript’s coding potential with the premise that a true protein coding gene is more likely to have a long and high-quality ORF. Different from CPC and CPAT, PhyloCSF is a method to determine whether a multi-species nucleotide sequence alignment is likely to represent a protein-coding region. PhyloCSF uses an algorithm that assumes that bona-fide protein coding genes are likely to contain high frequencies of synonymous codon substitutions and conservative amino acid substitutions but low frequencies of other missense and nonsense substitutions among orthologues. Experimentally, ribosome profiling is a technique that combines ribosome capture technology and high throughput sequencing to identify the location of translation start sites, their distribution, and the speed of the translating
ribosomes, thereby to differentiate non-protein coding genes from protein coding genes based on whether ribosome bind with RNA transcripts.

Functional study of lncRNA

Loss of function and gain of function studies are crucial in order to address the role of lncRNAs by screening the phenotypic change after knockdown or overexpression of lncRNA of interest in vitro or in vivo. There are several options for RNA interference. Among all options, siRNA (small interfering RNA) and shRNA (short hairpin RNA) are the most commonly used methods. siRNA is formed in the cell from shRNA or from dsRNA synthesized extracellular which cleaved by the Dicer enzyme. Compared to siRNA, shRNA is from nuclear expression and can persist in the cell for a long time. However, these oligonucleotide-based techniques share limitations: incomplete knockdown and unpredictable off-target effects. In order to avoid these limitations, gene-targeted knock-out technology provides powerful tool for functional study of lncRNAs in vivo. For example, CRISPR interference (CRISPRi) is a new genetic perturbation technique that allows for sequence-specific repression or activation of gene expression through the recruitment of activator or repressor to target lncRNA site by inactivated Cas9 system. In addition, the ectopic expression of the RNA may be able to mimic the function of the endogenous RNA. LncRNA overexpression requires the cloning of its sequence into an appropriate vector and the transfection of the vector into the cells.

Determination of lncRNA-protein interaction

LncRNAs have been shown to perform their functions in cooperation with protein. Therefore, the identification of lncRNAs interacting partner is crucial for characterizing functions of lncRNAs. In recent years, several new methodologies have been developed in order to obtain the lincRNA and protein or DNA interacting information. 1) RNA pulldown assay is a technique used to study the lncRNA:protein interaction. LncRNA binding protein can be identified using chemically modified extracellular synthesized RNA to pull down the RNA binding protein followed by western blotting or mass spectrometry. However, this technique has some limitations
including high false positive rate. 2) RAP (RNA antisense purification), ChIRP (chromatin isolation by RNA purification) and CHART (capture hybridization analysis of RNA targets) are all RNA-centric method. They are based on the same basic idea – using biotinylated oligonucleotides complementary to the lncRNA of interest to pull down associated proteins and chromatin. The only difference is that RAP uses 120 nucleotides overlapping biotinylated RNA probes tiled across the entire length of the target lncRNA, which enhance the specificity. Computationally, catRAPID is an algorithm to predict the interaction of protein-RNA pairs by combining secondary structure, hydrogen bond and van der Waals force.

THE STUDY OF LncRNA IN RAINBOW TROUT AND CATTLE

LncRNA study in rainbow trout

Rainbow trout (Oncorhynchus mykiss) is the most important aquatic animal species in the United States and all over the world. It belongs to the family of Salmonidae. In addition, rainbow trout is being used as a model in research in fields like carcinogenesis, immunology, toxicology and evolutionary biology. To improve aquaculture production and facilitate scientific study, a large amount of study and genetic information has been accumulated which include linkage map, physical map, and genome reference. However, complete understanding of the biological processes of rainbow trout is far from complete. Identifying lncRNA in rainbow trout would contribute to the current repertoire of rainbow trout lncRNA and genome annotation, and help further improve our understanding of the mechanism of biological processes and the evolutionary importance of these transcripts. Additionally, a recent study revealed that most of the important genetic marker, like SNP, located out of the protein-coding region where mutations could contribute to important economic traits. So far, three independent studies have discovered a number of lncRNAs expressed in zebrafish using RNA-Seq. The initial two studies revealed lncRNAs in early developmental stages of zebrafish with potential functions in embryogenesis. The latter study catalogued the lncRNA expression in five major tissue type of adult zebrafish. However, no lncRNA study has
been reported in rainbow trout.

The production of gonadal steroids increases as fish progress through sexual maturation\textsuperscript{66}. In fish the only period in the life cycle when estrogen levels rise substantially is during sexual maturation\textsuperscript{66}. The rise in concentration of estrogen in female rainbow trout induces hepatic production of vitellogenin, a precursor to yolk proteins\textsuperscript{67}. The process of sexual maturation requires high energy for the synthesis of the yolk in the egg, which is derived either exogenously or endogenously. Endogenous energy demands in fish are derived from the major protein source, skeletal muscle. In response of these energy demands, muscle undergoes rapid degradation, which eventually create an imbalance in the protein turnover during spawning or vitellogenesis\textsuperscript{67}. The effect of the sex steroids on protein turnover was well studied and confirmed that the imbalance of protein turnover is because of both increased catabolic pathways and decreased anabolic pathways\textsuperscript{67}, which results in skeletal muscle loss and poor quality muscle where the muscle protein is replaced with water.

Protein degradation occurs through three main routes of proteolysis including the ubiquitin-proteasome pathway, cathepsin-containing lysosomes, and calcium-dependent calpains\textsuperscript{67}. In addition, the caspase system stimulates proteolysis during apoptosis. IGF-1 is the signaling molecule, involved in the regeneration of muscle\textsuperscript{68}, activating the expression of different genes responsible for the synthesis of muscle. Previous study showed that elevation in plasma estrogen during maturation promotes protein catabolism via increased expression of proteolytic genes within multiple pathways of protein degradation, including ubiquitin-proteasome pathway, cathepsin-containing lysosomes and caspase system\textsuperscript{67}.

So far, a large range of functions has been attributed to lncRNAs, such as modulation of apoptosis and invasion\textsuperscript{69}, reprogramming of induced pluripotent stem cells\textsuperscript{70}, marker of cell fate\textsuperscript{71}, and parental imprinting\textsuperscript{72}. The steroid receptor RNA activator (SRA) is a lncRNA which was reported as coactivator of the estrogen receptor (ER) alpha in human\textsuperscript{73}. Moreover, the ability of lncRNAs to function as competing endogenous RNA (ceRNA) was first demonstrated in mouse and human myoblasts. Recent studies have shown that lncRNA controls muscle differentiation by functioning as a competing endogenous RNA (ceRNA)\textsuperscript{54} or by giving rise to microRNAs to promote
skeletal muscle differentiation and regeneration. Particularly, the ability of lncRNAs to function as ceRNA was first demonstrated in muscle differentiation. A muscle-specific lncRNA, linc-MD1, governs the time of muscle differentiation by sequestering miRNAs including miR-133 and miR-135 that specifically target protein coding genes MAML1 and MEF2C, which is needed to activate the differentiation program. Therefore, there is reason to speculate that lncRNAs might play important roles in muscle cell proliferation and differentiation. Analysis of the lncRNAs being expressed in muscle cells under the influence of estrogen will identify potential lncRNAs contributing to the muscle loss during vitellogenesis and spawning in rainbow trout. Furthermore, identification of lncRNAs allows for determining the responsive pathway contributing to the muscle loss and to further understand the role of the potential lncRNAs in muscle regeneration.

LncRNA study in cattle

Cattle (Bos taurus) is one of the most commonly raised livestock for meat, milk and other dairy products. There are a number of studies reporting bovine lncRNA across many tissues. Because of the key role of lncRNAs in regulation of gene expression, it is important to identify all lncRNAs in cattle. A total of 449 putative lncRNAs have been identified using public bovine expressed sequence tags sequences. Moreover, more than 4,000 lncRNAs were predicted in bovine skin using RNA-Seq data. Billery et al. (2014) identified a stringent set of 584 lincRNAs in bovine muscle. More recently, Koufariotis et al. (2015) reported a total of 9,778 lncRNAs from RNA-Seq data across 18 tissues. However, very little is known about the identity and characteristics of lncRNAs in bovine oocytes. The developmental competence of an oocyte, also known as egg quality, is defined as the ability of the egg to be fertilized and subsequently develop into a normal embryo. Mammalian oocytes become transcriptionally silent following germinal-vesicle breakdown, so the final stages of oocyte maturation and early embryo development depend on stored transcripts. Several studies have reported that lncRNA plays critical roles in the embryonic stem cell regulatory. For instance, more than 100 lincRNA promoters were identified to be bound by stem cell factors such as OCT4 and Nanog. Therefore, the study of lncRNA in bovine oocyte could help understand the embryonic development.
REFERENCES


Brockdorff, N. *et al.* The product of the mouse Xist gene is a 15 kb inactive X-


Cesana, M. *et al.* A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* 147, 358-369,


Kaushik, K. et al. Dynamic expression of long non-coding RNAs (lincRNAs) in


CHAPTER 1
Identification and functional prediction of large intergenic noncoding RNAs (lincRNAs) in rainbow trout (*Oncorhynchus mykiss*)

Jian Wang¹, Liyuan Fu¹, Prasanthi P. Koganti¹, Lei Wang¹, Jacqelyn M. Hand¹, Hao Ma² and Jianbo Yao¹*

¹Division of Animal and Nutritional Sciences, West Virginia University, Morgantown, WV 26506 USA

²USDA/ARS National Center for Cool and Cold Water Aquaculture, Kearneysville, WV, 25430 USA

*Corresponding author
*Published in Marine Biotechnology. doi: 10.1007/s10126-016-9689-5.
ABSTRACT

Long noncoding RNAs (lncRNAs) have been recognized in recent years as key regulators of diverse cellular processes. Genome-wide large-scale projects have uncovered thousands of lncRNAs in many model organisms. Large intergenic noncoding RNAs (lincRNAs) are lncRNAs that are transcribed from intergenic regions of genomes. To date, no lincRNAs in non-model teleost fish have been reported. In this report, we present the first reference catalog of 9,674 rainbow trout lincRNAs based on analysis of RNA-Seq data from 15 tissues. Systematic analysis revealed that lincRNAs in rainbow trout share many characteristics with those in other mammalian species. They are shorter, lower in exon number and expression level compared with protein-coding genes. They show tissue-specific expression pattern and are typically co-expressed with their neighboring genes. Co-expression network analysis suggested that many lincRNAs are associated with immune response, muscle differentiation and neural development. The study provides an opportunity for future experimental and computational studies to uncover the functions of lincRNAs in rainbow trout.

KEYWORDS: Co-expression network, LncRNA, LincRNA, Rainbow trout
INTRODUCTION

Long noncoding RNAs (lncRNAs) are non-coding RNA molecules that are longer than 200 nucleotides and carry many signatures of mRNAs, such as 5' capping, 3' polyadenylation, and RNA splicing, but have little or no open reading frame \(^1-3\). They have emerged as a new class of regulatory transcripts in recent years \(^4,5\). Recent advances in sequencing technologies have opened a new horizon for the identification and annotation of this class of RNAs in many species. The lncRNAs that are transcribed from intergenic regions of genomes are termed large intergenic noncoding RNAs (lincRNAs). As lincRNAs do not overlap with protein-coding regions, computational analysis of such RNAs is easier. To date, at least 15,512 human lincRNAs and over 10,000 mouse lincRNAs have been identified \(^6,7\).

Recent studies have supported the view that lincRNAs play important roles in many biological processes, such as procession of p53 response pathways \(^8-10\), regulation of epigenetic marks and gene expression \(^11-14\), maintenance of pluripotency \(^15\) and activation of gene expression as “enhancer RNAs” \(^16,17\). In addition, lincRNAs have also been associated with human diseases and pathophysiological conditions \(^18-20\).

Rainbow trout (Oncorhynchus mykiss) is a species of salmonid native to cold-water tributaries of the Pacific Ocean in Asia and North America. It is one of the most important cold water fish species in the USA due to its importance for food production, sport fisheries, and its importance for environmental carcinogenesis research \(^21\). To generate genomic resources for genetic studies of this species, we have characterized the rainbow trout mRNA and microRNA transcriptomes \(^22-25\). In particular, a complete transcriptome has been generated by RNA-sequencing of cDNA libraries from multiple tissues of a single doubled haploid rainbow trout \(^24\). With the increasing evidence supporting important roles of lincRNAs in diverse processes, a systematic catalog of these RNA transcripts and their expression across tissues in rainbow trout is warranted. The recent publication of rainbow trout genome sequence \(^26\) and computational methods for transcriptome reconstruction \(^27-29\) provide an opportunity to comprehensively annotate and characterize lincRNA transcripts in rainbow trout.

Here we report the systematic identification and characterization of lincRNAs in 15
major tissue types of rainbow trout. We analyzed the known genomic features of the identified lincRNAs including transcript length, exon number and spatiotemporal expression specificity. We also used weighted gene co-expression network to assign functionalities to the lincRNAs, which revealed that lincRNAs are expressed in a strong tissue-specific manner, and many of them are highly associated with biological processes specific to that tissue (e.g., a brain-specific group is enriched with functional terms such as neural development and axon injury response). This study is the first report of a genome-wide annotation of rainbow trout lincRNAs, which will facilitate future experimental and computational studies to uncover the functions of lincRNAs in rainbow trout.
MATERIALS AND METHODS

Tissue sample collection and RNA sequencing

Tissue collection and RNA sequencing were described in detail in a previous study \(^{24}\). In brief, 13 different tissues were collected from a single male homozygous rainbow trout, which was euthanized under protocol #02456 approved by the Washington State University Institutional Animal Care and Use Committee. These tissues include brain, fat, gill, head kidney, intestine, kidney, liver, testis, red muscle, skin, spleen, stomach and white muscle. In addition, oocyte and pineal samples were collected from different fish. Total RNA from each sample was isolated using Trizol (Invitrogen, Carlsbad, CA). Library construction and sequencing were performed at Roy J. Carver Biotechnology Center, University of Illinois at Urbana-Champaign. Each library was loaded onto one lane and paired-end sequencing with 2x100 cycles was performed on an Illumina Genome Analyzer IIx (Illumina, San Diego, CA).

RNA-Seq reads mapping and transcriptome assembly

Spliced read aligner TopHat version V2.0 \(^{28}\) was used to map all sequence reads to the rainbow trout genome \(^{26}\). A two-step mapping process was performed by TopHat using the following parameters: min-anchor = 5, min-isoform-fraction = 0, and default values for the remaining parameters. Bowtie2 \(^{30}\) was used first to align reads with no gaps that can directly map to the genome reference sequence. Gapped alignment was then performed to align the reads that were not aligned in the first step. Aligned reads from each sample were assembled into transcriptome by Cufflinks version V2.2.1 \(^{31}\) that uses spliced reads information to determine exons connectivity. The Cufflinks assembler generates the output in the form of FPKM (Fragments Per Kilobase of exons per Million fragments generated) value, which is directly proportional to the relative abundance of a transcript in a given sample.

FPKM threshold for classifying complete and partial transcripts

Individual transcript assembly may have noise from multiple sources such as artifacts generated by sequence alignment, unspliced intronic pre-mRNA or genomic
DNA contamination. Sebnif\textsuperscript{32}, an integrative bioinformatics pipeline that identifies high-quality single- and multi-exonic lincRNAs by optimizing a FPKM threshold, was used to minimize the assembly noise and enhance the quality of identified lincRNAs. Considering the difference of the structure between the multi- and single-exonic transcripts, 2 separate algorithms were used to identify the optimal FPKM thresholds. 1) For multi-exonic transcripts, a Fully Reconstruction Fraction Estimation (FRFE) approach was used by Sebnif\textsuperscript{27}. Briefly, multi-exonic transcripts in reference annotation were first divided into N expression quantiles based on their FPKM values. At each expression quantile, the reference transcript set was then divided into 2 categories, fully reconstructed transcripts and partially reconstructed transcripts. The assembly quality was evaluated by the proportion of the fully reconstructed transcripts, which is also called Fully Reconstruction Fraction (FRF), at each expression quantile. The index of the optimum FPKM threshold was obtained by balancing the sensitivity and specificity based on the FRF value with the following formula\textsuperscript{33}.

\[ i^* = \arg \min_{i \in I} \sqrt{\left(1 - \text{sensitivities}[i]\right)^2 + \left(1 - \text{specificities}[i]\right)^2} \]

Where \(i^*\) is the index of FPKM threshold for each quantile \(i\). The sensitivity \([i]\) and specificity \([i]\) indicate the \(i\)th sensitivities and specificities, respectively. The \(i\) belongs to \([1, N]\). The optimum FPKM threshold was generated by pROC\textsuperscript{34}. 2) For single-exonic transcripts, Single-exonic Transcript Gaussian/Gamma Estimation (STGE), was implemented to estimate the optimal expression threshold\textsuperscript{32}. In the STGE algorithm, the appropriate model was determined by fitting the expression values of the single-exonic transcripts in the reference annotation. Any transcript whose expression falls into either tail of fitting model distribution was considered unreliable and discarded.

LincRNA detection pipeline

A step-wise filtering pipeline (Figure 1) was used to identify putative lincRNAs from deep sequencing data. 1) All transcripts smaller than 200 bases were excluded. 2) Assembled transcripts were annotated using cuffcompare from cufflinks\textsuperscript{31}. Transcripts that are located in the intergenic region, at least 1 kb away from any known protein-coding genes, were selected as putative lincRNAs\textsuperscript{7}. 3) Coding potential of each
transcript was calculated using CPAT (Coding-Potential Assessment Tool)\(^\text{35}\) and CPC (Coding Potential Calculator)\(^\text{36}\). 4) To evaluate which of the remaining transcripts contains a known protein coding domain, HMMER-3\(^\text{37}\) was used to identify transcripts translated in all six possible frames having homologs with any of the 31,912 known protein family domains in the Pfam database (release 24; both PfamA and PfamB). All transcripts with a Pfam hit were excluded based on the E-value cutoff of 0.01. 5) Putative protein-coding RNAs were filtered out by applying a maximal ORF (Open reading frame) length threshold. Any transcripts with a maximal ORF > 100 amino acids was excluded. 6) Sequence homology search was performed to remove those transcripts with significant similarity with RNAs in several different public RNA databases including Rfam\(^\text{38}\), RNAdb\(^\text{39}\) and lncRNAdb\(^\text{40}\). 7) The remaining transcripts located at least 1 kb from any known protein-coding genes were selected\(^\text{7}\).

Tissue specificity score and neighboring gene correlation analysis

To evaluate tissue specificity of a transcript, an entropy-based metric that relies on Jensen-Shannon (JS) divergence was used to calculate specificity scores (0 to 1). A perfect tissue-specific pattern is scored as JS=1, which means a transcript is expressed only in one tissue\(^\text{41}\). In neighboring gene analysis, 2 genes were defined as neighbors if the minimal distance between them is < 10 kb (regardless of their directions)\(^\text{7,42}\). The expression correlation between 2 neighbors was estimated by calculating the Pearson correlation coefficient between their density-normalized expression values (log2 FPKM + 1).

Weighted gene co-expression network construction and gene module detection

All genes with expression variance ranked in the top 75 percentile of the data set were retained\(^\text{2}\). R package “WGCNA” was then used to construct the weighted gene co-expression network\(^\text{43}\). A matrix of signed Pearson correlation between all gene pairs was computed, and the transformed matrix (TOM) was used as input for linkage hierarchical clustering\(^\text{43}\). Genes with similar expression patterns were clustered together.
Functional enrichment analysis

To investigate the potential roles of lincRNAs in rainbow trout, we performed Blast2GO analysis to assign gene ontology (GO) terms to all protein-coding genes associated with lincRNAs in each network module. A cutoff value of 1E-10 was used for BLASTx. GO terms enrichment analysis were performed using Fisher’s exact test (p-value <0.01). Interaction networks among lincRNA and protein-coding genes were constructed based on co-expression using Cytoscape (http://www.cytoscape.org/).

Validation of expression specificity of lincRNAs

Expression specificity of selected lincRNAs was validated by RT-PCR analysis as described previously. PCR primers are listed in Supplemental file 1. Tissue samples used in the analysis include brain, fat, gill, head kidney, intestine, kidney, liver, testis, red muscle, skin, spleen, stomach, white muscle, oocyte and pineal. 18S rRNA was used as a control for RNA quality.
RESULTS AND DISCUSSION

Transcriptome reconstruction and filtering low-quality assemblies

To comprehensively identify rainbow trout lincRNAs, we collected and deeply sequenced the RNA samples from brain, fat, gill, head kidney, intestine, kidney, liver, testis, red muscle, skin, spleen, stomach, white muscle, oocyte and pineal gland. A total of 1.3 billion raw paired-end sequence reads (100-bp read length) were generated from 15 rainbow trout tissues. The number of reads from each tissue ranged from 78.8 to 93.5 million. A total of 1,087,497,866 cleaned reads (81.4%) were harvested for further analysis. These sequence reads were mapped to the rainbow trout genome using TopHat and approximately 447 million (82%) mapped reads were recovered. The mapping ratio ranged from 76.9% to 89.5% with an average of 82.3% (Table 1). We then used the ab initio assembly software Cufflinks to reconstruct the transcriptome for each tissue based on the read-mapping results (Figure 1A). On average, 79,021 transcripts for each tissue were obtained.

The first challenge to annotate lincRNA gene loci is to distinguish lowly expressed lincRNAs from the tens of thousands of lowly expressed unreliable fragments assembled from RNA-Seq. To address this challenge, we removed unreliable lowly expressed transcripts using a learned FPKM threshold, which was calculated using Sebnif (Figure 1B). First, we classified all transcripts that did not overlap the genomic region of a known protein-coding gene annotation of rainbow trout as novel intergenic transcripts (category of ‘u’ assigned by cuffcompare) and defined an average of 28,012 ‘u’ transcripts for each tissue (Figure 1B, Supplemental file 2), among which 6,975 and 21,037 are multi- and single-exonic transcripts, respectively. Next, FRFE and STGE algorithms were used to distinguish partial transcripts from full-length transcripts. For 6,975 multi-exonic transcripts, sebnif applied a FRFE threshold of 0.5. For 21,037 single-exonic transcripts, STGE was used to model the transcript expression profiles with the lower and upper probability cutoffs set at 0.05 and 0.95, respectively. Following this filtering, an average of 4,628 multi-exonic (FPKM > 2.76) and 4,071 single-exonic (FPKM >3.14) transcripts for each tissue were retained. Finally, a total of 39,745
intergenic transcripts were obtained by merging all intergenic transcripts from 15 tissues.

Identification and characterization of rainbow trout lincRNAs

The currently available coding potential prediction methods only work well for protein coding RNAs. Therefore, the most widely used strategy to annotate potential non-coding RNAs (ncRNA) is to exclude those that possess protein coding features. The filtering pipeline we used to identify novel lincRNAs is shown in Figure 1C. First, we analyzed the coding potential of unannotated transcripts using CPAT and CPC, which filtered out 61% (24,329) of all transcripts. Second, we scanned each transcript in all six frames to exclude transcripts that contain any of the 31,912 protein coding domains cataloged in the protein family database Pfam. This filtering retained 10,773 potential lincRNA transcripts. Furthermore, a minimal ORF length criterion was applied to distinguish lincRNAs from mRNAs. A cutoff of 300 nt (100 codons) was used to exclude putative mRNAs. For characterization of ncRNAs not yet annotated in the rainbow trout genome assembly, sequence homology search was performed to exclude those transcripts with significant similarity with RNAs in Rfam, RNAdb and lncRNAdb. Finally, we identified 9,674 lincRNAs after removing those transcripts that are located within 1 kb from any known protein-coding genes (Supplemental file 3).

Previous studies in mammals have shown that lncRNAs are shorter, less conserved, and expressed at significantly lower level compared with protein-coding genes. To determine whether rainbow trout lincRNAs have similar features, we characterized the basic features of the identified lincRNAs by comparing them with protein-coding genes. We found that rainbow trout lincRNAs are on average about half of the length of protein-coding genes (mean length of 705 nt for lincRNAs vs. 1,635 nt for protein coding transcripts) (Figure 2A). Moreover, lincRNAs had fewer exons (on average, 1.3 exons for lincRNAs vs. 6.9 exons for protein-coding genes) (Figure 2B). Notably, the mean length and average exon number of rainbow trout lincRNAs are smaller than those of human (~1000 nt and 2.9 exons) and zebrafish (~1000 nt and 2.8 exons). This could be due to underestimation of the length and exon number of rainbow trout lincRNAs resulting from their lower abundance and lower sequencing depth (incomplete assembly). Furthermore, the expression levels of lincRNAs are on average about 10-fold lower than those of
Rainbow trout lincRNAs show more tissue-specific expression pattern than protein-coding genes

Recent studies have shown that lincRNAs are expressed in a more tissue-specific manner than protein-coding genes\textsuperscript{41}. We analyzed expression pattern for each lincRNA transcript. Of the 9,674 potential lincRNAs, 8,545 were expressed in more than one tissue (Figure 4A, 4B). The remaining 1,129 lincRNAs displayed tissue-specific expression (Figure 4D). Among the 15 tissues, brain expressed the most number of tissue-specific lincRNAs (161), which is consistent with the result from a previous study in zebrafish\textsuperscript{50}. Skin, white muscle and liver had relatively lower numbers of tissue-specific lincRNAs (Figure 4C). The tissue specificity score for each lincRNA was calculated using an entropy-based metric that relies on Jensen-Shannon (JS) divergence\textsuperscript{41}. Results showed that 46% of rainbow trout lincRNAs were tissue-specific, relative to only 18% of protein-coding genes (Figure 5). Thus, rainbow trout lincRNAs exhibited more tissue-specificity than protein-coding genes (P < 10\textsuperscript{-16}, Fisher exact test), which is in agreement with data from other species\textsuperscript{27,41,49}.

Tissue-specific expression of lincRNAs determined by computational analysis was validated by RT-PCR analysis. A total of 10 lincRNAs were selected for validation of their expression in 15 tissues. They included 7 lincRNAs specifically expressed in a particular tissue (Linc-OM9284 in brain, Linc-OM8822 in red muscle, Linc-OM8901 in intestine, Linc-OM3900 in stomach, Linc-OM8614 in testis, Linc-OM8334 in fat, Linc-OM8318 in kidney), 2 lincRNAs expressed in 2 tissues (Linc-OM8912 in oocyte and skin, Linc-OM9283 in skin and liver), and 1 lincRNA ubiquitously expressed in all tissues (Linc-OM9274). As shown in Figure 6, the RT-PCR result matches perfectly with the expression profiles estimated from deep sequencing data.
Co-expression of lincRNAs with neighboring coding genes

The occurrence of pairs of neighboring lincRNA: protein-coding genes within expression clusters suggests that such organization may be important for the regulatory function of lincRNAs. Recent studies indicated that some lincRNAs may act in cis and regulate the expression of genes in their chromosomal neighborhood. One expectation of the cis hypothesis is that the expression of lincRNAs and their neighboring genes would be correlated across all tissue samples. Therefore, we analyzed the expression patterns of 1,146 (12%) of identified lincRNAs that are located within 10 kb from a coding gene. We observed a more correlated expression pattern of lincRNAs and their neighboring coding genes (mean correlation: 0.211) compared to random coding gene pairs (mean correlation: 0.042) (P < 2×10^{-16}, Kolomogorv-Smirnov [KS] test) (Figure 7). Meanwhile, lincRNAs: coding gene pairs also exhibited a modestly higher correlative expression pattern than coding gene pairs (mean correlation: 0.115) (P < 2.2×10^{-16}, KS test). On the other hand, there is also a significant difference between neighboring coding gene pairs and random coding gene pairs (P < 7.9×10^{-13}, KS test). This observation suggests that the correlation between lincRNAs and their neighbor coding genes is higher than both neighboring coding gene pairs and random coding gene pairs.

Functional prediction of lincRNAs based on co-expression network

The comprehensive lincRNA catalog allows us to investigate the potential functions of these novel transcripts in rainbow trout. Here, we built a co-expression network to associate lincRNAs with mRNAs by performing weighted gene co-expression network analysis (WGCNA) and inferred the putative lincRNA functions based on “guilty-by-association” analysis. By clustering correlated genes together, we identified 34 co-expression gene modules containing 2,963 lincRNAs and 10,321 protein-coding genes in total. Notably, 6 of 34 modules are related to immune response, muscle differentiation and neural development based on the enriched GO terms associated with their modules (Figure 8).

The functional annotations enriched in 4 modules (blue, grey60, tan and green) are functionally related to immune responses (Figure 8B and Supplemental file 5). In each of
these 4 modules, we observed many lincRNAs that are highly expressed in spleen, gill, and intestine (Figure 8A), suggesting that these lincRNAs might be involved in immune related processes. In blue module, many genes were enriched in T cell receptor signaling and PI3K/AKT/mTOR signaling pathways. The lincRNAs that are co-expressed with tyrosine-protein kinase (ITK), which phosphorylates PLCγ1 in T cell signaling 52, may play important roles in T cell signaling and function. PI3K and mTOR signaling pathways are important in regulating immune cell activation in neutrophils and mast cells, and type I interferon production 53. Those lincRNAs that are co-expressed with PI3K or mTOR pathway genes are likely involved in these immune processes. In grey module, the lincRNAs that are co-expressed with integrin, which mediates immune cells to penetrate into tissues 54, may play critical roles in immune cell migration and cell-cell interactions that occur during the course of an immune response. In the module, lincRNAs that are co-expressed with Rab20, a key player in phagosome maturation 55, may function in phagocytosis. Likewise, lincRNAs in green module are co-expressed with MHC class I genes 56, indicating that they might be involved in processing and presenting antigen to T cells.

Cyan module contains transcripts (165 protein coding genes and 15 lincRNAs) that are highly expressed in muscle (Figure 8). Most of the enriched genes in this module are related to functions or development of muscle (Supplemental file 5). Notably, the lincRNAs that are co-expressed with myoblast determination protein 2 (MyoD2) may play roles in regulating muscle differentiation. A previous study has demonstrated the role of a specific lncRNA in controlling muscle differentiation 57.

Recent studies have shown that many lincRNAs are brain-specific, indicating a role in brain development 58,59. This study also found that brain had the most tissue-specific lincRNAs (Figure 4C). The lincRNAs in lightyellow module are co-expressed with genes important for neural differentiation and development, such as dihydropyrimidinase-related protein (DRP) and Draxin precursor, indicating that they may function as important regulators of neurogenesis.

Collectively, the functional prediction analysis revealed that tissue-specific lincRNAs and protein-coding genes are enriched for processes specific to that tissue and essential in maintaining each tissue’s identity and functionality.
CONCLUSION

In this report, we provided the first comprehensive annotation of rainbow trout lincRNAs based on whole transcriptome sequencing of multiple tissues, and identified 9,674 novel lincRNA transcripts. These lincRNAs tend to be expressed in tissue-specific manner, and share many characteristics with those in mammalian species. Co-expression network analysis suggested many rainbow trout lincRNAs are associated with immune response, muscle differentiation and neural development. The study lays the groundwork for future functional characterization of lincRNAs in rainbow trout.
Acknowledgements

This study was supported by the USDA ARS Cooperative Agreement No. 58-1930-0-059. It is published with the approval of the director of the West Virginia Agricultural and Forestry Experiment Station as scientific paper No. 3268.
REFERENCES


32 Sun, K., Zhao, Y., Wang, H. & Sun, H. Sebnif: an integrated bioinformatics


Weichhart, T. & Saemmann, M. D. The PI3K/Akt/mTOR pathway in innate


Figure 1. Pipeline used to identify novel lincRNAs. (A) Raw RNA-Seq data was pre-processed and mapped using Tophat, and assembled using Cufflinks in ab initio mode. (B) Sebnif was used to filter all lowly expressed unreliable transcripts. (C) Pipeline for lincRNA detection.
Figure 2. Structural characteristics of lincRNAs in comparison to protein-coding transcripts. (A) Transcript length. Shown are cumulative distributions of transcript length for the lincRNAs (red), protein-coding genes (blue). Data is shown after elimination the coding genes which are larger than 8KB. (B) Analysis of lincRNA exon number. The distribution of exon number for different sets of lincRNAs and coding genes is shown with color as in (A). All coding genes with more than 20 exons was eliminated.
Figure 3. Comparison of the expression levels of lincRNAs and protein-coding genes.
Maximal expression abundance (log2-normalized FPKM counts as estimated by Cufflinks) of each lincRNA (red) and protein coding genes (green).
Figure 4. Tissue-wise distribution of predicted novel lincRNAs.

(A) Distribution of 9674 potential lincRNAs across 15 tissues. Figure 5A depicts the number of lincRNAs that were expressed either in single or multiple tissues. (B) Venn diagram representing 7783 lincRNAs across five tissues viz; Gill (blue-A), Intestine (yellow-B), Kidney (orange-C), Spleen (green-D), Stomach (pink-E). (C) The distribution of tissue specific lincRNAs across 15 tissues. (D) The heat map of 1129 tissue specific lincRNAs across 15 tissues. Each individual heat map represents the expression level in the parent tissue versus other tissues based on FPKM values.
Figure 5. Tissue specificity of lincRNAs and protein-coding genes.
Shown are distribution of maximal tissue specificity scores calculated for each transcript across all tissues.
Figure 6. Validation of expression specificity of lincRNAs by RT-PCR analysis.

Expression of 10 selected lincRNAs was analyzed by RT-PCR in rainbow trout tissues including brain (Br), oocyte (Oo), white muscle (Wm), pineal (Pi), fat (Fa), gill (Gi), skin (Sk), head kidney (Hk), testis (Te), spleen (Sp), stomach (St), liver (Li), red muscle (Rm), intestine (In) and kidney (Ki). 18S rRNA was used as a control for RNA quality.
Figure 7. Correlation of expression patterns between pairs of neighboring genes.

Shown are distributions of Pearson correlation coefficients in expression levels across the tissues between 1,146 pairs of lincRNAs and their neighboring coding genes (green), 9,363 pairs of coding gene neighbors (blue), and 8,000 random pairs of protein-coding genes (red).
Figure 8. Function prediction of rainbow trout lincRNAs.

(A) Upper panel: heatmaps showing expression patterns of all genes in each co-expression gene modules across 15 tissues. Middle panel: bar plots showing the corresponding module eigengene expression value. Lower panel: pie charts showing ratio of mRNAs and lincRNAs in each module. (B) Functional enrichment in each module. Length of bars indicates the significance (-log10 transformed FDR).
Table 1. Summary of samples and RNA-Seq data.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Reads</th>
<th>Clean reads</th>
<th>Mapped reads</th>
<th>Mapping ration (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain</td>
<td>84,816,430</td>
<td>72,852,798</td>
<td>58,713,834</td>
<td>80.6</td>
</tr>
<tr>
<td>Fat</td>
<td>93,546,068</td>
<td>77,569,466</td>
<td>64,131,248</td>
<td>82.7</td>
</tr>
<tr>
<td>Gill</td>
<td>92,670,670</td>
<td>77,800,922</td>
<td>62,957,778</td>
<td>80.9</td>
</tr>
<tr>
<td>Head</td>
<td>92,168,818</td>
<td>77,800,922</td>
<td>62,957,778</td>
<td>80.9</td>
</tr>
<tr>
<td>Kidney</td>
<td>92,613,688</td>
<td>65,824,376</td>
<td>54,975,550</td>
<td>83.5</td>
</tr>
<tr>
<td>Intestine</td>
<td>89,642,288</td>
<td>73,508,012</td>
<td>60,173,822</td>
<td>81.9</td>
</tr>
<tr>
<td>Liver</td>
<td>85,281,910</td>
<td>67,792,890</td>
<td>55,075,712</td>
<td>81.2</td>
</tr>
<tr>
<td>Oocyte</td>
<td>90,135,204</td>
<td>73,938,296</td>
<td>60,609,542</td>
<td>82.0</td>
</tr>
<tr>
<td>Red muscle</td>
<td>93,064,168</td>
<td>70,234,780</td>
<td>56,954,940</td>
<td>81.1</td>
</tr>
<tr>
<td>Skin</td>
<td>87,743,778</td>
<td>65,230,154</td>
<td>55,286,532</td>
<td>84.8</td>
</tr>
<tr>
<td>Spleen</td>
<td>93,532,200</td>
<td>74,312,778</td>
<td>57,535,030</td>
<td>77.4</td>
</tr>
<tr>
<td>Stomach</td>
<td>91,231,186</td>
<td>77,577,260</td>
<td>64,002,482</td>
<td>82.5</td>
</tr>
<tr>
<td>White Muscle</td>
<td>86,643,770</td>
<td>72,910,820</td>
<td>56,083,986</td>
<td>76.9</td>
</tr>
<tr>
<td>Pineal</td>
<td>87,802,668</td>
<td>76,591,074</td>
<td>63,895,584</td>
<td>83.4</td>
</tr>
<tr>
<td>Testis</td>
<td>85,389,746</td>
<td>66,567,414</td>
<td>57,260,414</td>
<td>86.0</td>
</tr>
<tr>
<td>Total</td>
<td>1,336,282,592</td>
<td>1,087,497,866</td>
<td>894,601,592</td>
<td>82.3</td>
</tr>
</tbody>
</table>
CHAPTER 2
Comprehensive analysis of lncRNAs and mRNAs in E2 treated skeletal muscle of rainbow trout (Oncorhynchus mykiss)

Jian Wang¹§, Prasanthi P. Koganti¹§, Jianbo Yao¹ and Beth Cleveland²*

¹Division of Animal and Nutritional Sciences, West Virginia University, Morgantown, WV 26506 USA

²USDA/ARS National Center for Cool and Cold Water Aquaculture, Kearneysville, WV, 25430 USA

*Corresponding author
§Co-first author
ABSTRACT

Estrogen (E2) is a steroid hormone that negatively affects muscle growth and protein homeostasis in rainbow trout, but the mechanisms associated with this response are not fully understood. To better characterize the effects of E2 in muscle, we identified differentially regulated mRNAs and lncRNAs in juvenile rainbow trout exposed to E2. Here, we performed next-generation RNA sequencing and comprehensive bioinformatics analyses to characterize the transcriptome profiles, including mRNAs and long noncoding RNAs (lncRNAs), in skeletal muscle of normal and E2 treated rainbow trout. A total of 226 lncRNAs and 253 mRNAs were identified as differentially regulated. We identified crucial pathways, including several signal transduction pathways, hormone response, oxidative response and protein, carbon and fatty acid metabolism pathways. Subsequently, a functional lncRNA-mRNA co-expression network was constructed, which consisted of 681 co-expression relationships between 164 lncRNAs and 201 mRNAs. Moreover, an lncRNA-pathway network was constructed. A total of 65 key lncRNAs were identified, which regulate 20 significantly enriched pathways. Finally, the function of a novel lncRNA (lnc-OM9822) was predicted, which may be activated by estrogen receptor alpha (ER1) and involve in the estrogen-signaling pathway. Overall, our analysis not only effectively provides insights into the mRNA and lncRNA association with the effect of E2 on skeletal muscle in rainbow trout but also provides further insights into understanding the molecular mechanism of lncRNAs.
INTRODUCTION

Cell survival and its response to various environmental cues are the outcomes of functional proteins. Inherited information is passed from DNA to RNA and then proteins as indicated by the principle of central dogma. For decades we have been trying to understand the necessary factors contributing to gene expression and its regulation. Advances in this area presented various levels of gene regulation during the processes of transcription of mRNA, translation of active protein and their modifications. Identification of interacting molecules that regulate gene expression is necessary.

Rainbow trout is an important fresh water fish and known for its importance as a research model. Reproductive phases in this salmonid is crucial as the energy expenditure to address the synthesis and release of oocytes is taxing. Skeletal muscle during the reproductive phase acts as an endogenous source to address the energy demand, compromising muscle quality$^{1,2}$. Reduced muscle quality in turn results in reduced fillet quality. Reproduction in female fish starts with the release of steroid hormone estrogen initiating synthesis of vitellogenin in liver$^{3}$. Vitellogenin is an important oocyte protein necessary for the synthesis of eggs. Estrogen is released only during reproductive phases in fish. Immediate effects of estrogen on skeletal muscle and its gene expression are not clearly understood. Myogenic precursor cells (MPCs) proliferate and differentiate to form mature muscle cells contributing to hyperplastic and hypertrophic growth in salmonids. Muscle synthesis involves different signaling pathways, myogenic regulatory factors (MRFs) and epigenetic factors and their interactions. Different stage specific MRFs play a role in this process, of which MyoD is an important MRF necessary for the differentiation of MPCs to myoblasts.

Estrogen receptor (ER) is stimulated by its binding ligand, estradiol. Cytoplasmic and membrane bound ERs function either through genomic or non-genomic means. Genomic activation involves the activation of estrogen receptors in the presence of estrogen, and their binding either to the estrogen response elements or the other transcription factors to activate gene expression. Estrogen receptors in non-genomic mechanism activate different signaling pathways in target tissues, including MAPK$^{4-6}$, Phosphoinositol 3-kinase$^{7}$ pathways. This mechanism involves the activation of different
scaffold proteins \(^8\) and proximal signaling proteins \(^9,10\). The signaling pathways initiate cascade of reactions by phosphorylating different signaling molecules thereby activating the transcription factors, ultimately regulating gene expression. The current study focuses on not only understanding differentially expressed genes in skeletal muscle exposed to estrogen but also their interaction with long non-coding RNAs.

Availability of sequencing techniques led to the discovery of various kinds of non-coding transcripts involved in regulation of gene expression along with protein coding genes. Evidences from these studies indicate that more than 50\% of the transcripts belong to non-coding RNA species. Transcripts with lengths more than 200 bp and no protein coding potential are long non coding RNA (lncRNA)\(^{11}\). LncRNA have fewer, longer exons when compared to coding genes and exhibit cell-type specific expression. They interact with DNA, RNA and proteins to regulate gene expression at transcriptional, post-transcriptional, chromatin remodeling and epigenetic levels\(^{12}\). The role of lncRNA in skeletal muscle development is relatively well studied in humans and mice. Various studies supported their role in transcriptional regulation of skeletal muscle development and differentiation\(^{12}\). For instance, ChiP-Seq studies led to the identification of two different long noncoding RNAs (\(^{\text{DRR}}\)RNA and \(^{\text{CE}}\)RNA) that transcribe from the enhancer region of myogenic transcription factor MyoD\(^{12,13}\). These long noncoding RNAs were proved to enhance the expression of MyoD and myogenin. Similar studies identified a number of muscle specific lnc RNAs regulated by Yin Yang 1 protein and are called YY1 associated lncRNA\(^{14}\). They function either as positive or negative regulators of muscle differentiation.

To date, different studies highlight the roles of lnc RNAs or miRNAs in skeletal muscle development; however, emphasis on their interaction and with protein coding genes is the focus of our present study. The complex myogenesis network is not only regulated by individual functional non-coding and coding RNAs but also by their interactions. The present study aims to identify the differentially expressed lncRNAs and mRNAs in skeletal muscle exposed to E2 and create a network of the mRNAs and lncRNAs involved in the process. This is the first of its kind of study in farm animals.

In this study, we used RNA-Seq approach to comprehensively investigate the genome-wide gene expression differences in skeletal muscle under the effect of E2. We
also constructed a functional lncRNA-mRNA regulatory network that is associated with skeletal muscle response to E2 treatment, and conducted a pathway enrichment analysis. Moreover, we found one lncRNA with largest degree in the co-expression network which contains an estrogen response element in an exon region and can interact with protein of estrogen receptor alpha (ER1). This comprehensive analysis helped to understand the effect of E2 on fish skeletal muscle and provided novel insights into the roles of lncRNA in protein turnover under the effect of E2 at transcriptomic level.
RESULTS

Generation of a muscle transcriptome reference

To provide comprehensive understanding of the effects of E2 treatment on protein turnover in rainbow trout, the skeletal muscle samples were collected from E2 treated and control fish at two time points, 24 hours and 72 hours after injections and subsequently sent for deep sequencing. A total of 789,485,036 paired-end raw reads were generated from 16 samples with 101-bp read length (4 biological replicates of each treatment at respective time points). The number of sequences from each sample ranged from 36.2 to 63.8 million. After quality control including removal of ambiguous nucleotides, low-quality bases and ribosomal RNA sequences, a total of 749,490,934 cleaned reads (94.9%) were harvested for further analysis. The number of cleaned reads of each sample ranged from 34.5 to 60.9 million (Table 1).

The cleaned reads of 16 samples were pooled and assembled by Trinity\textsuperscript{15}. Then CD-HIT-EST\textsuperscript{16} was used to remove the redundancy. As shown in Table 2, the transcriptome were assembled into 243,509 contigs (203,148 Trinity components), ranging from 201 to 20,635 bp in length. To provide reliable reference and filter out transcripts that had very low read counts, The R package EdgeR\textsuperscript{17} was used to remove the transcripts with count-per-million (CPM) less than 1 in any of the four replicate for each sample. As the result, 63,181 contigs (31,419 Trinity components) were generated, ranging from 201 to 20,635 bp in length. The average length is 1,466 bp, N50 length is 1,982 bp and median length is 1,189 bp (Table 2). All cleaned reads were mapped to the reliably expressed transcriptome reference by using the ultrafast short read aligner Bowtie\textsuperscript{18}. The mapping ratio ranged from 84.3% to 89.8% with an average of 87.7% (Table 1).

Identification of differentially expressed genes

The pipeline reported in our previous study\textsuperscript{19} was used to identify all IncRNAs from the reliably expressed transcriptome reference. Two biological replicates (CTRL3 & EST4) at 24 hour time point were excluded for the exact test according to multidimensional distance scaling analysis (Supplementary Fig. S1). Finally, differential expression analyses were performed of coding genes and IncRNAs in E2 treated fish at
24 hours and 72 hours compared with the respective control fish. Given a false discovery rate (FDR) of 5%, 479 (226 lncRNA and 253 mRNA) differentially expressed genes (DEGs) were obtained between E2 and control fish at 24 hours and 19 (9 lncRNA and 10 mRNA) DEGs were obtained at 72 hours. In these DEGs, only 1 lncRNA and 1 mRNA were detected differentially expressed at both time points (Fig. 1A, 1B). These results suggest that the effects of E2 after 72 hours is limited in skeletal muscles. Therefore, further analysis was performed with the differentially expressed mRNAs and lncRNAs with skeletal muscle samples after 24 hours of injection.

In the E2 treated group at time point of 24 hours, 102 up-regulated and 151 down-regulated mRNAs and 119 up-regulated and 117 down-regulated lncRNAs were identified compared to the control group (Fig. 1C,D,E). As shown in Fig. 1D and Fig. 1E, heat maps were generated using the differentially expressed lncRNAs and mRNAs respectively and they were clearly self-segregated into control (CTRL) and EST (E2 treated) clusters in unsupervised hierarchical clustering analysis (Fig. 1D,E). There were 36 genes with more than 16-folds difference between control and E2 treated sample, of which 14 were up-regulated in EST and 12 of those were down-regulated. These results reflect the influence of E2 on altering the gene expression in muscle of rainbow trout. Additionally, only three out of 226 differentially expressed lncRNAs (Linc-OM206, 3272 &7282) were previously reported in rainbow trout lncRNA reference identified from 15 tissues including muscle, which indicated that lncRNAs, as reported, are more temporospatial expressed, and that a well annotated rainbow trout lncRNA reference has yet to be achieved.

**Experimental validation of lncRNA and mRNA**

To confirm the sequencing results, seven lnc RNAs and five mRNAs that showed significant differential expression were randomly selected for real time PCR validation. All the selected RNA showed similar trends with the expression profiles in vivo. Fig. 2 A, B represents the log2fold change from sequencing and real time PCR of lncRNA and mRNA respectively. P-values obtained from the t-test performed for the qPCR results were represented depending on their significance.
Differentially expressed genes represent several important pathways

To obtain the functions of the differentially expressed mRNAs and connections among them, we performed GO term and KEGG pathway enrichment analyses. To better interpret the resulting enriched GO terms, all redundant GO terms were merged with the representative terms. In GO analysis of the differentially expressed mRNAs between E2 treated and control group, the most significantly over-represented terms are shown in Table 3, including the following biological process (BP) terms: peptide cross-linking, response to testosterone, mast cell migration, neurotransmitter biosynthetic process, positive regulation of receptor biosynthetic process, mitochondrial transport, female pregnancy, catechol-containing compound biosynthetic process, activation of JAK2 kinase activity and cellular response to leptin stimulus in the up-regulated genes, and mesenchyme morphogenesis, mesenchyme migration, cAMP biosynthetic process, gluconeogenesis, tissue migration, cristae formation, enteroendocrine cell differentiation and bicarbonate transport in down-regulated genes; molecular function terms: protein-glutamine gamma-glutamyltransferase activity, hormone binding, hormone receptor binding, transferring amino-acyl activity and tyrosine 3-monooxygenase activity in up-regulated genes, and actin binding, cytoskeletal protein binding, motor activity, calcium ion binding and ATPase inhibitor activity in down-regulated genes; cellular component terms: sex chromosome, ubiquitin conjugating enzyme complex and MHC protein complex in up-regulated genes, and actin cytoskeleton, troponin complex, cytoskeleton, contractile fiber and intracellular non-membrane-bounded organelle in down-regulated genes.

To have a general view of the whole picture and comprehensively interpret the GO term enrichment result, treemap was constructed based on the result of semantic similarity analysis. In each treemap, those loosely related terms were joined into a ‘supercluster’ with the same color and the most significant term as the representative of the group. As shown in Fig. 3A,B, in response to E2, skeletal muscle tissue showed notable up-regulations mostly related to oxidative stress response, hormone response, protein ubiquitination, cysteine biosynthesis and DNA repair, and down-regulations mostly related to mesenchyme morphogenesis and cAMP biosynthesis in biological process. Notably, the term of response to leptin (GO:0044321) was identified to be
enriched (Supplementary Fig. S2). Up-regulated genes in molecular function mostly related to kinase activity, hormone receptor activity, amino-acyl transferase activity and ubiquitin conjugating enzyme activity, and of down-regulated genes mostly related to muscle structural protein binding, motor activity, pyruvate carboxykinase activity and glucose-6-phosphate activity (Fig. 3C,D). The enrichment of up-regulated genes in cellular component mostly related to chromosome and ubiquitin conjugating enzyme complex, and of down-regulated genes mostly related to skeletal muscle structural protein complex (Fig. 3E,F).

The KEGG pathway analyses mapped the differentially expressed genes to KEGG reference pathways to infer systemic biological behaviors. Comparing to the control group, we observed significant KEGG pathway enrichment of differentially expressed genes in skeletal muscle in response to E2 stress. Of the top 20 over-represented KEGG pathways (Fig. 4), 4 were involved in signal transduction, including Rap1 signaling pathway, Jak-STAT signaling pathway, calcium signaling pathway and AMPK signaling pathway. Additionally, 4 of the top enriched KEGG pathways were associated with endocrine system, which are PPAR signaling pathway, insulin resistance, estrogen signaling pathway and adipocytokine signaling pathway. Finally, the remaining 12 over-represented KEGG pathways were classified into different functional groups, including amino acid and carbohydrate metabolism, cellular community, protein synthesis and digestion.

**Construction of IncRNA-mRNA co-expression network**

To construct the differentially expressed IncRNA-mRNA co-expression network, the normalized expression values of the IncRNAs and mRNAs were obtained by using DESeq2. Subsequently, Pearson’s correlation coefficient (PCC) was calculated between the normalized expression values of each of the IncRNA-mRNA pairs. The IncRNA-mRNA pairs with a PCC >0.99 and FDR<0.05 were selected for network construction. Finally, the co-expression network was constructed, which consisted of 681 co-expression relationships between 164 IncRNAs and 201 mRNAs (Fig. 5). We then considered the node degree of the network, as a higher degree indicated that the nodes were likely to be hubs and therefore involved in more competing interactions. In the
present study, as a more stringent threshold, the top 5% (top 20) of the nodes were defined as hubs (Table 4). These 20 hub nodes that contain 14 lncRNAs co-expressed with more than 40% of the nodes in the network, implying the centrality of these nodes.

To better understand the possible function of lncRNA, lncRNA-pathway network was constructed, in which nodes represented lncRNAs or over-represented pathways. They were connected if the co-expressed mRNAs of a lncRNA were involved in the corresponding over-represented pathway, indicating that these pathways were regulated by the corresponding lncRNAs (Fig. 6, Supplementary Table S1, S2). Finally, 65 lncRNAs were linked with 20 significantly enriched pathways in the lncRNA-pathway network, suggesting the key-regulating role of these lncRNAs in response to E2 stress. The key lncRNAs and corresponding crucial pathways are shown in Fig. 6 and in Table S3 and S4.

ESR1 related lncRNA (Inc-OM9822).

In pathway enrichment analysis, the estrogen signaling pathway was shown to be over-represented (Fig. 4) as expected with estrogen treatment. Of the 3 differentially expressed genes involved in estrogen signaling pathway (Fig. 7B), one gene encodes estrogen receptor alpha (ESR1) which is a ligand-activated transcription factor activated by the sex hormone estrogen. ESR1 is composed of several domains which are crucial for hormone binding, DNA binding and activation of transcription. Moreover, in co-expression networks, Inc-OM9822 with highest connectivity was strongly connected to ESR1 and estrogen signaling pathway (Fig. 7A). To further understand if this lncRNA interacts with estrogen receptors, Comet was used to analyze the presence of estrogen response elements. One estrogen response element was found to be a target site for the estrogen receptor (Fig. 7C). Moreover, catRAPID and lncPRO were both used to analyze the interaction between ESR1 and Inc-OM9822. Two interactions between ESR1 and Inc-OM9822 were found (Fig. 7D). These results help predict the relationship between the lncRNA and estrogen receptor. Studies to clearly understand the regulation are necessary. Additional motif searching was performed using MEME software to examine if Inc-OM9822 has any repetitive sequences which is an important character of lnc RNA. The search was performed to identify 5 motifs across the sequence. Three of the motifs
sequences ranging from 21 to 29 bp (Fig. 8A and B). Motif 1 and motif 2 are highly repetitive relative to the rest with 4 and 5 copies respectively.
DISCUSSION

The present study led to the identification of lncRNAs and mRNAs regulated by estrogen in rainbow trout skeletal muscle. RNA sequencing was performed to understand the inhibited or stimulated expression of lncRNAs and mRNAs by E2. Gene ontology analysis reveals the role of coding genes as cellular components and in different biological processes and molecular functions. Further coexpression analysis was performed to understand the potential functions of lncRNAs.

Up-regulated representative GO terms regulated by E2

Products of various genes functioning as molecular factors, cellular components and those involved in biological processes were differentially expressed in skeletal muscle under the influence of estrogen. They function to regulate different signaling pathways including JAK/STAT pathway, response to oxidative stress, cell-cell signaling and intracellular estrogen receptor signaling pathways with positive regulation of various receptor biosynthesis and their binding activity (insulin receptor binding, insulin like growth factor binding, integrin binding etc). Besides, important protein cross-linking molecules involved in histone ubiquitination, protein autoubiquitination, and their responsible enzymes like ubiquitin conjugating enzyme and its complex were also up-regulated. Elevated expression of important metabolic enzymes including aldehyde dehydrogenase and succinate dehydrogenase important for energy synthesis was observed.

Estrogen (E2) hormone induces the activation of estrogen receptors that are functional through genomic and non-genomic pathways. Besides activation of estrogen receptors evidences for the role of E2 in regulation of IGF pathway and JAK/STAT pathway. E2 is released during reproductive phases in salmonids contributing to the disruption of GH/IGF-1 axis as well as decrease in IGF-1. IGF binding proteins are specific proteins that carry IGF to their receptors as well regulate the availability of IGF. Six different binding proteins were identified in rainbow trout IGFBP-1 to IGFBP-6. IGFBP-5 was reported to be abundantly expressed in skeletal muscle during myoblast differentiation. Interestingly, IGFBP-5 was reported to inhibits muscle cell differentiation by binding to the IGF receptor and impeding IGF activity in studies using...
mouse embryonic fibroblasts\textsuperscript{29}. IGFBP-5 is differentially expression in skeletal muscle treated with E2 within 24 hours observed in our study. Real time expression analysis of IFGBP-5 after 24 hours of treatment showed four-fold increase in estrogen treated skeletal muscle compared to control. Collectively these observations indicate that an increase in IGFBP-5 in skeletal muscle treated with E2 might act as a balancing factor for muscle cell differentiation.

Evidences for E2 regulation of JAK/STAT pathway and their interplay at the transcriptional level through coactivators using CHIP experiments were reported\textsuperscript{23}. JAK/STAT pathway was known to be involved in proliferation and/or differentiation of different cells including osteoblasts, myoblasts and immune response cells\textsuperscript{30-33}. JAKs belong to the non-receptor tyrosine kinase receptor family activated by ligands followed by STATs. Nature of different JAKs and STATs and their function in various pathways were well studied in mammals\textsuperscript{34}. All four JAKs identified in mammals and six of the seven orthologs of STATs were found in teleosts\textsuperscript{35}. The JAK/STAT pathway is also involved in regulation of estrogen receptors expression as a result of E2 response. For example, elevated expression of JAK2 was observed in hormone responsive tissues with prolonged exposure to E2\textsuperscript{36} and is reported to be required for the proliferation of MPCs\textsuperscript{37}. The pathway plays a prominent role in the process of myogenesis and is involved from proliferation, differentiation to terminal differentiation besides interacting with different myogenic factors and signaling pathways responsible for myogenesis. Interactions with myogenic factors, MyoD and MEF2\textsuperscript{38} and with different signaling pathways including IGF-AKT, HGF-ERK\textsuperscript{39} were reported. Studies with C2C12 and mouse models disclosed the role of JAK1-STAT1-STAT3 in activation of myogenic precursor cell (MPCs) proliferation and inhibition of their differentiation\textsuperscript{40,41}. JAK2-STAT3 is partially responsible for the proliferation of MPCs\textsuperscript{42}. In contrast JAK2-STAT2-STAT3 promotes differentiation of MPCs\textsuperscript{39}. A negative role of JAK3 in terminal differentiation was reported as well\textsuperscript{43}. STATs are also the regulatory factors of MPCs proliferation and differentiation. STAT1 associates with MEF2 and contributes to the inhibition of differentiation while STAT3 regulates both proliferation and differentiation positively\textsuperscript{40-42}. 
Studies in Javanese medaka (*Oryzias javanicus*), Japanese sea bass (*Lateolabrax japonicus*) and laboratory animals under continuous exposure to E2 resulted in production of reactive oxygen species leading to oxidative stress. Most of the stress related genes analyzed in Javanese medaka under the exposure to estrogen showed an increased and a dose depended expression in skeletal muscle. Reactive oxygen species interact with DNA causing damage and such damage was reported in fish exposed to estrogen. Reactive oxygen species (ROS) released as products of oxidative stress is one among the other contributors of autophagy. Autophagy in skeletal muscle is due to an increase in ROS and its activity through ULK and MAPK. Besides autophagy ROS also contributes to muscle atrophy by activating muscle degrading pathways involving calpains, caspase-3, ATG4b, MuRF-1 and atrogin-1. Studies to understand the miRNA and their target genes with the same samples used in the present experiment showed significant increases in the expression of caspase-3 and atrogin-1 in samples treated with estrogen at 24 hours (Koganti data not published). Myostatin, belonging to TGF-β family is a well known negative regulator of myogenesis. Myostatin regulates muscle atrophy by inducing the production of ROS and signaling through NF-kB pathway. *In vivo* and *in vitro* experiments with rainbow trout fish and myocytes showed that exposure to E2 resulted in increase of ubiquitin ligase genes resulting in increased proteolytic pathways in fish. Atrogin-1 is an ubiquitin ligase responsible to add ubiquitin to the target proteins intended for degradation by proteasome. MyoD and myogenin are the target molecules of atrogin-1, as indicated by studied in mammals. Increased expression of ubiquitin ligases in skeletal muscle under the effect of E2 was previously reported in rainbow trout (Koganti, manuscript in preparation). Real time PCR in the present study showed an increased expression of autophagy related 4b cysteine peptidase (ATG4b) expression in skeletal muscle samples under the influence of estrogen. Conclusively, these results indicate the regulation of E2 on the skeletal myogenesis and its protein turnover.

**Down-regulated representative GO terms regulated by E2**

Characterized GO terms that showed reduced expression in estrogen treated skeletal muscle after 24 hours are mainly involved in structural make up and of skeletal
muscle. Significantly down regulated proteins primarily are involved in actin cytoskeleton, actin binding, motor activity and mesenchymal morphogenesis. Additionally, GO terms related to cAMP biosynthesis, cristae formation, regulation of smoothened signaling pathway, RNA mediated DNA polymerase activity were differentially regulated. Mesenchymal morphogenesis contributes to embryonic development and adult development. Their migration and morphogenesis is necessary for adult tissue turnover. Mesenchymal stromal cells are multipotent cells that differentiate to different cell lineages including mesodermal cells considered as precursors for myoblasts or MPCs\(^{55,56}\). Adult muscle regeneration involves the activation of the muscle precursor cells, which are specific for the muscle cell lineage. They are present between the basal lamina and sarcolemma of the mature muscle bundle\(^5\). On injury or need of regeneration they proliferate and differentiate to form a mature muscle fiber which contributes to the muscle mass. There is a need for these MPCs to replenish. The bone marrow derived mesenchymal cells with myogenic markers migrate to the site of muscle regeneration\(^{58,59}\). Decrease in the genes responsible for mesenchymal cells migration and morphogenesis as observed in the present study results in depletion of MPCs thus creating an imbalance in protein turnover and eventually poor quality muscle. Activation of myogenesis in bone marrow derived cells is through Wnt signaling\(^6\). Experiments with rat bone marrow derived mesenchymal stem cells supplemented with Wnt3a showed that Wnt signaling is necessary for the differentiation of mesenchymal cells to myogenic origin as well as their migration\(^6\). Wnt3a not only promotes cells to myogenic origin but also inhibits their differentiation to adipogenic origin by decreasing the expression of CCAAT enhancer binding protein alpha and peroxisome proliferator-activated receptor gamma (PPAR\(\gamma\))\(^6\). PPAR\(\gamma\) is a transcription factor responsible in regulating expression of genes involved in energy and lipid metabolism\(^6\). Peroxisome proliferator-activated receptor gamma coactivator 1 alpha (PPARGC1A) is a coactivator of this nuclear transcription factor which is reported to be highly expressed in skeletal muscle\(^6,63\). PPARGC1A is expressed mainly in type I fibers and are also responsible for mitochondrial biogenesis and oxidative metabolism\(^6,65\). PPARGC1A is also reported to be a principle regulator of muscle fiber type\(^63\) and increase intramyocellular lipid accumulation\(^66\). Studies also indicate that PPARGC1A and PPARGC1B play a prominent role in estrogen receptor
signaling pathway. They act as cofactors and enhances transactivation of estrogen receptor $\alpha^{67-69}$. Expression analysis of PPARGC1A in skeletal muscle treated with E2 showed an increase at 24 hours compared to control. These results indicate that E2 significantly alter the oxidative metabolism and fiber type in skeletal muscle.

Actin is a conserved protein highly expressed in striated muscles and is involved in various cellular functions such as maintenance of cell shape, receptor mediated cell response, regulation of transport (ionic and motor based)$^{70}$. De novo formation of actin filaments is called nucleation. Polymerization of the nucleated actin filaments occurs at their barbed ends. Similarly actin molecules are removed in the same way. Continuous assembly and disassembly of actin filaments occur in any cell$^{70,71}$. Large number of actin binding proteins involving in various functions were identified$^{72}$. Actin binding proteins function as monomer binding proteins, filament-depolymerizing proteins, filament end-binding proteins, filament severing proteins, cross-linking proteins, stabilizing proteins and motor proteins. Contractile function in a cell is due to the actin-based motors belonging to myosin family. Downregulation of these proteins result in changes in actin behavior in the cell resulting in changes in cell integrity.

LncRNAs are emerging as regulators of diverse biological functions$^{73}$. The significant functional molecular mechanism of lncRNAs has been continuously recognized, particularly in development and diseases$^{74}$. However, at present, only a few of lncRNAs were functionally well documented, and there is a lack of comprehensive databases that provide with the resource of experimentally verified lncRNA functions. Bioinformatics approach, such as co-expression analysis, was mainly used to infer lncRNA functions, in which lncRNA functions were predicted based on the functions of their co-expressed genes, as genes that exhibit similar expression patterns under multiple conditions have a tendency to be involved in the same pathways$^{75}$. Co-expression models have been performed by integrating the expression profiles of protein-coding genes and lncRNAs for large-scale prediction of lncRNA functions in a coding-non-coding gene co-expression network$^{76,77}$. Therefore, our work expanded knowledge of lncRNA associated protein turnover process under effect of E2 and paly an important role as a pre-processing step to guide further lab experimental designs.
In the present study, we investigated the expression patterns of lncRNAs and mRNAs and constructed a functional lncRNA-mRNA regulatory network for E2 treated skeletal muscle protein turnover change in rainbow trout. As the result, we found that 226 lncRNAs were differentially expressed, indicating that lncRNAs may be associated with unbalanced protein turnover process. We hypothesized that both differentially expressed lncRNAs and mRNAs were associated with the unbalance of protein turnover under effect of E2. Furthermore, the co-expression relationship between these differentially expressed lncRNAs and mRNAs was investigated, and a co-expression network was constructed, which contain a total of 164 lncRNAs and 201 mRNAs. This network can provide a global view of all possible lncRNA-coding gene expression associations based on the E2 treatment background. In this network, top 5% (20) large degree nodes were all lncRNAs. Of which, 14 lncRNAs co-expressed with more than 40% of the nodes. We hypothesized that although not coding for a protein, these lncRNAs might be involved in the unbalance of skeletal muscle protein turnover after treatment with E2.

To examine the key lncRNAs and their potential functions, lncRNA-pathway network was constructed based on pathway enrichment analysis. The lncRNA-pathway network contains 65 lncRNAs linked with 20 significantly enriched pathways. The top 1 pathway with most lncRNAs linked is focal adhesion. Focal adhesions are integrin-containing, multi-protein structures that mediate the regulatory effects of a cell in response to extracellular matrix (ECM) adhesion\textsuperscript{78}. Focal adhesion was extensively studied in tumor cell. It was reported that E2 induce cell migration through activation of focal adhesion kinase (FAK) in endometrial cancers\textsuperscript{79}. In addition to regulation of migration, the researchers found that focal adhesions (FAs) also contribute to ECM degradation\textsuperscript{80}. More importantly, disrupting the FAs complex significantly impairs FAK-mediated degradation\textsuperscript{80}. These results suggest that focal adhesion related lncRNAs might involve in the skeletal muscle degradation under the effect of E2. We also found that calcium signaling pathway was linked with 15 lncRNAs in our lncRNA-pathway network. Calcium ions are important for cellular signaling, which plays a pivotal role in almost all cellular processes. Calpain is a protein belonging to the family of non-lysosomal cysteine proteases, which is activated by calcium ions. Calpains are believed to function in various biological processes, including apoptosis\textsuperscript{81}. Calsequestrin, a down-
regulated gene in our study, is a calcium-binding protein which regulate calcium homeostasis in cell. The pathways with large degree also include those pathways related to carbon and fatty acid metabolism and important signal transduction. Although the results of the present study require further experimental verification, the results provide further insight into understanding the role of lncRNAs in skeletal muscle degradation under effect of E2.

Identification of Inc-OM9822 in co-expression studies as well as its differential expression in 24 hour samples explains its prominent role in skeletal muscle under the influence of estrogen. This long noncoding RNA expression is influenced by E2 treatment at 24 hours. Presence of an estrogen receptor element (ERE) in Inc-9822 indicate binding of estrogen receptor with Inc-OM9822. Long noncoding RNA are proved to be functional in many different ways including functional interaction with different proteins. LncRNA are proved to interact with different transcription factors and either prevent their association with DNA or act as coactivator to enhance gene expression\(^{82,83}\). Such lncRNA are called trans-regulators\(^{84}\). For instance, lncRNA Gas5 folds mimicking the DNA that binds to the transcriptional regulator, glucocorticoid receptor thus regulating its availability and activity\(^{85}\). Contrasting to the function of Gas5, lncRNA SRA was discovered to function as a coactivator of MyoD which is a transcription factor necessary for myoblast differentiation\(^{83}\). Long noncoding with repeat regions are known to play a prominent role in post-transcriptional as well as translational regulation of gene expression\(^{82,86}\), emphasizing the functional roles of repeat regions present in lncRNAs. Higher degree correlation of Inc-OM9822 revealed its interaction with various signaling pathways, most of them having proven to have direct or indirect effect on skeletal muscle synthesis. This study helps us understand the influence of E2 on such signaling pathways while focusing on the role of lncRNA as an additional layer of gene regulation.

In summary, our comprehensive analyses provided novel knowledge of mRNAs and lncRNAs at the transcriptomic level during the influence of E2 on rainbow trout skeletal muscle within 24 hours. These results and conclusions may serve as important resources for future experimental dissections of lncRNAs in rainbow trout.
METHODS

Ethics statement

All animal experiments in this study were performed at the USDA/ARS National Center for Cool and Cold Water Aquaculture (NCCWCA) according to protocol #50 approved by the NCCWCA Institutional Animal Care and Use Committee.

Experimental design

A total of twenty fish weighing approximately 40 g were sampled and divided into two groups randomly with ten fish in each group for each treatment (10 fish per treatment). The study consisted of two treatments, including intraperitoneal injections of E2 and the delivery vehicle to serve as the control. E2 was resuspended (10 µg/µL) in 95% ethanol and diluted to 2.5 µg/µL with vegetable oil. The control treatment contained an equal ratio of ethanol: vegetable oil as compared to E2 suspension. E2 and the vehicle injection methodology was adapted from previously published procedures used in tilapia. Feed was withheld the day of E2 injection and throughout the study period. Fish were anesthetized with tricaine methanesulphonate (MS-222, 100 mg/l), weighed, and received intraperitoneal injections (2.0 µl/g body weight) of E2 (5.0 µg/g body weight) or the vehicle. Skeletal muscle was dissected and collected 24 and 72 hours post-injection, immediately frozen using liquid nitrogen for further processing.

cDNA library construction and sequencing

Total RNA was extracted from skeletal muscle samples of fish treated with E2 and controls at 24 and 72 hours (6 biological replicates each) using TRIZOL reagent (Invitrogen, Carlsbad, CA). Quality and quantity of RNA was estimated using the $A_{260}:A_{280}$ ratio. Integrity and size distribution were evaluated on Bioanalyzer 2100 (Agilent technologies, Santa Clara, CA). Four replicates of each treatment were sent for sequencing.

RNA extraction and quality control

Total RNA was extracted from skeletal muscle samples of fish treated with E2 and controls at 24 and 72 hours (6 biological replicates each) using TRIZOL reagent (Invitrogen, Carlsbad, CA).
Sequence data processing, de novo assembly and differential expression analysis

Adaptor sequences were trimmed and ambiguous and low quality bases were removed. Then read length less than 50 were removed. TRINITY\textsuperscript{15} was used to assembly all cleaned reads with default parameters. CD-HIT-EST was used to remove the shorter redundant transcripts when they were 100\% covered by other transcripts with more than 99\% identity\textsuperscript{16}. All the cleaned reads were mapped to the assembled transcriptome by Bowtie\textsuperscript{18}. RSEM was used to estimate and quantify gene expression levels from RNA-Seq data\textsuperscript{88}. The final counts matrix file was used as input for the R package edgeR\textsuperscript{17} to normalize the data, and to filter out genes that can’t possibly be expressed in all the samples for any of the conditions (CPM>1) to generate reference for differential gene expression analysis. The exact test in edgeR was carried out to discover the DEGs between the groups in different experimental conditions. FDR was used for multiple test correction. Any genes with fold change >2 and FDR <0.05 were defined as DEG.

Validation of sequencing results

Sequencing results were validated using real time PCR. RNA from 24 hour samples ( 6 replicates from each treatment) were used for cDNA synthesis using miScript II (Qiagen, Valencia, CA). Expression of mRNA and IncRNA was normalized to the endogenous control, β-actin gene. Real time PCR was performed following the standard procedure including melt curve analysis. Relative expression was calculated by comparing the expression of treatment samples to control samples. T-test was performed to determine statistical differences in gene expression.

LncRNAs identification and GO and KEGG enrichment analysis of mRNAs

All DEGs were mapped to rainbow trout genome using blat\textsuperscript{89}. The pipeline reported in our previous study\textsuperscript{19} for IncRNA identification was followed to detect differentially expressed IncRNAs in present study. All differentially expressed mRNAs were subjected to similarity search against NCBI non-redundant (nr) protein database using BLASTx\textsuperscript{90} with an e-value cutoff of 1e-10. Gene names and GI were assigned to each mRNA based on the BLASTx result. ID mapping was performed using our in house script to extract all associated GO id for each mRNA. KEGG pathways were assigned to each mRNA using the online KEGG Automatic Annotation Server (KASS)\textsuperscript{91}. The R
package GOstats was used to run hypergeometric testing on GO and KEGG terms\textsuperscript{92}. Redundant GO terms were removed by REVIGO, a Web server that summarizes long, unintelligible lists of GO terms by finding a representative subset of terms using semantic similarity measurement based clustering algorithm\textsuperscript{93}.

**Co-expression analysis**

To identify co-expressed lncRNA-mRNA pairs, Pearson’s correlation coefficients were calculated based on the normalized expression value between every differentially expressed lncRNA and mRNA pair. Only the strong correlations (0.99 or greater) were selected to construct the network. The threshold of FDR was set to <0.05. Cytoscape was used to construct the co-expression network\textsuperscript{94}. Finally, 681 co-expression relationships between 164 lncRNAs and 201 mRNAs were identified.
REFERENCES


Wang, J. *et al.* Identification and Functional Prediction of Large Intergenic Noncoding RNAs (lincRNAs) in Rainbow Trout (Oncorhynchus mykiss). *Marine


27 Kamangar, B. B., Gabillard, J. C. & Bobe, J. Insulin-like growth factor-binding protein (IGFBP)-1, -2, -3, -4, -5, and -6 and IGFBP-related protein 1 during


Summermatter, S., Baum, O., Santos, G., Hoppeler, H. & Handschin, C. Peroxisome Proliferator-activated Receptor γ Coactivator 1α (PGC-1α) Promotes Skeletal Muscle Lipid Refueling in Vivo by Activating de Novo Lipogenesis and the Pentose Phosphate Pathway. *Journal of Biological Chemistry* **285**, 32793-


88 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq


<table>
<thead>
<tr>
<th>Group</th>
<th>Time Point (h)</th>
<th>Replicate</th>
<th>Reads</th>
<th>Clean reads</th>
<th>Mapped reads</th>
<th>Mapping ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTRL</td>
<td>24</td>
<td>1</td>
<td>36,802,826</td>
<td>35,432,594</td>
<td>31,145,250</td>
<td>87.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>61,291,464</td>
<td>52,228,282</td>
<td>46,274,258</td>
<td>88.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>59,111,132</td>
<td>56,868,138</td>
<td>50,328,302</td>
<td>88.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>50,601,976</td>
<td>48,340,122</td>
<td>42,974,368</td>
<td>88.9</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>1</td>
<td>63,019,266</td>
<td>60,702,946</td>
<td>54,389,838</td>
<td>89.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>60,452,644</td>
<td>58,114,660</td>
<td>52,186,986</td>
<td>89.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>51,642,380</td>
<td>49,378,116</td>
<td>44,045,278</td>
<td>89.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>63,821,958</td>
<td>60,981,886</td>
<td>53,786,022</td>
<td>88.2</td>
</tr>
<tr>
<td>EST</td>
<td>24</td>
<td>1</td>
<td>61,743,118</td>
<td>59,403,804</td>
<td>52,394,128</td>
<td>88.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>37,513,122</td>
<td>35,786,668</td>
<td>30,186,188</td>
<td>84.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>42,203,620</td>
<td>40,341,512</td>
<td>34,895,386</td>
<td>86.5</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>4</td>
<td>47,655,922</td>
<td>45,563,586</td>
<td>40,232,838</td>
<td>88.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>36,221,406</td>
<td>34,809,018</td>
<td>30,144,800</td>
<td>86.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>41,421,664</td>
<td>39,634,688</td>
<td>34,878,528</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>38,926,574</td>
<td>37,367,184</td>
<td>31,874,208</td>
<td>85.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>37,055,964</td>
<td>34,537,730</td>
<td>29,586,296</td>
<td>85.6</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>789,485,036</td>
<td>749,490,934</td>
<td>659,322,674</td>
<td></td>
</tr>
</tbody>
</table>

CTRL indicates control group. EST represents estrogen treatment group.
Table 2. Statistics of transcriptome assembly.

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Number of components</th>
<th>203,148</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of contigs</td>
<td>243,509</td>
</tr>
<tr>
<td></td>
<td>Maximum contig length</td>
<td>20,635 bp</td>
</tr>
<tr>
<td></td>
<td>Minimum contig length</td>
<td>201 bp</td>
</tr>
<tr>
<td></td>
<td>Average contig length</td>
<td>660 bp</td>
</tr>
<tr>
<td></td>
<td>Median contig length</td>
<td>365 bp</td>
</tr>
<tr>
<td></td>
<td>N50 length</td>
<td>1,076 bp</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Filtered Assembly</th>
<th>Number of components</th>
<th>31,419</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of contigs</td>
<td>63,181</td>
</tr>
<tr>
<td></td>
<td>Maximum contig length</td>
<td>20,635 bp</td>
</tr>
<tr>
<td></td>
<td>Minimum contig length</td>
<td>201 bp</td>
</tr>
<tr>
<td></td>
<td>Average contig length</td>
<td>1,466 bp</td>
</tr>
<tr>
<td></td>
<td>Median contig length</td>
<td>1,189 bp</td>
</tr>
<tr>
<td></td>
<td>N50 length</td>
<td>1,982 bp</td>
</tr>
</tbody>
</table>
Table 3. Most enriched representative GO terms of up- or down-regulated genes response to E2.

<table>
<thead>
<tr>
<th>Catagoray</th>
<th>GO term</th>
<th>Definition</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BP</strong></td>
<td>UP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0018149</td>
<td>peptide cross-linking</td>
<td>0.0003</td>
<td></td>
</tr>
<tr>
<td>GO:0033574</td>
<td>response to testosterone</td>
<td>0.0024</td>
<td></td>
</tr>
<tr>
<td>GO:0097531</td>
<td>mast cell migration</td>
<td>0.0024</td>
<td></td>
</tr>
<tr>
<td>GO:0042136</td>
<td>neurotransmitter biosynthetic process</td>
<td>0.0024</td>
<td></td>
</tr>
<tr>
<td>GO:0010870</td>
<td>biosynthetic process</td>
<td>0.0024</td>
<td></td>
</tr>
<tr>
<td>GO:0006839</td>
<td>mitochondrial transport</td>
<td>0.0035</td>
<td></td>
</tr>
<tr>
<td>GO:0007565</td>
<td>female pregnancy</td>
<td>0.0049</td>
<td></td>
</tr>
<tr>
<td>GO:0009713</td>
<td>catechol-containing compound biosynthetic process</td>
<td>0.0049</td>
<td></td>
</tr>
<tr>
<td>GO:0042977</td>
<td>activation of JAK2 kinase activity</td>
<td>0.0049</td>
<td></td>
</tr>
<tr>
<td>GO:0044320</td>
<td>cellular response to leptin stimulus</td>
<td>0.0049</td>
<td></td>
</tr>
<tr>
<td><strong>Down</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0072132</td>
<td>mesenchyme morphogenesis</td>
<td>0.0002</td>
<td></td>
</tr>
<tr>
<td>GO:0090131</td>
<td>mesenchyme migration</td>
<td>0.0002</td>
<td></td>
</tr>
<tr>
<td>GO:0006171</td>
<td>cAMP biosynthetic process</td>
<td>0.0047</td>
<td></td>
</tr>
<tr>
<td>GO:0006094</td>
<td>gluconeogenesis</td>
<td>0.0051</td>
<td></td>
</tr>
<tr>
<td>GO:0090130</td>
<td>tissue migration</td>
<td>0.0071</td>
<td></td>
</tr>
<tr>
<td>GO:0042407</td>
<td>cristae formation</td>
<td>0.0079</td>
<td></td>
</tr>
<tr>
<td>GO:0035883</td>
<td>enteroendocrine cell differentiation</td>
<td>0.0079</td>
<td></td>
</tr>
<tr>
<td>GO:0015701</td>
<td>bicarbonate transport</td>
<td>0.0119</td>
<td></td>
</tr>
<tr>
<td><strong>MF</strong></td>
<td>UP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:003810</td>
<td>protein-glutamine gamma-glutamyltransferase activity</td>
<td>0.0003</td>
<td></td>
</tr>
<tr>
<td>GO:0042562</td>
<td>hormone binding</td>
<td>0.0005</td>
<td></td>
</tr>
<tr>
<td>GO:0051427</td>
<td>hormone receptor binding transferase activity, transferring amino-acyl groups</td>
<td>0.0007</td>
<td></td>
</tr>
<tr>
<td>GO:0016755</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0004511</td>
<td>tyrosine 3-monooxygenase activity</td>
<td>0.0024</td>
<td></td>
</tr>
<tr>
<td><strong>CC</strong></td>
<td>UP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO:0003779</td>
<td>actin binding</td>
<td>1.44012E-06</td>
<td></td>
</tr>
<tr>
<td>GO:0008092</td>
<td>cytoskeletal protein binding</td>
<td>4.6302E-05</td>
<td></td>
</tr>
<tr>
<td>GO:0003774</td>
<td>motor activity</td>
<td>0.0002</td>
<td></td>
</tr>
<tr>
<td>GO:0005509</td>
<td>calcium ion binding</td>
<td>0.0032</td>
<td></td>
</tr>
<tr>
<td>GO:0042030</td>
<td>ATPase inhibitor activity</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>GO:000803</td>
<td>sex chromosome</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>GO:0031371</td>
<td>ubiquitin conjugating enzyme complex</td>
<td>0.008</td>
<td></td>
</tr>
<tr>
<td>GO:0042611</td>
<td>MHC protein complex</td>
<td>0.035</td>
<td></td>
</tr>
<tr>
<td>GO ID</td>
<td>Term</td>
<td>P-value</td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------------------------------</td>
<td>---------------</td>
<td></td>
</tr>
<tr>
<td>GO:0015629</td>
<td>actin cytoskeleton</td>
<td>1.38995E-09</td>
<td></td>
</tr>
<tr>
<td>GO:0005861</td>
<td>troponin complex</td>
<td>2.86022E-07</td>
<td></td>
</tr>
<tr>
<td>GO:0005856</td>
<td>cytoskeleton</td>
<td>6.84069E-07</td>
<td></td>
</tr>
<tr>
<td>GO:0043292</td>
<td>contractile fiber</td>
<td>3.95003E-06</td>
<td></td>
</tr>
<tr>
<td></td>
<td>intracellular non-membrane-bound organelle</td>
<td>4.35011E-05</td>
<td></td>
</tr>
</tbody>
</table>

BP indicates biological process. MF represents molecular function. CC indicates cellular component.
Table 4. Top 20 nodes with the highest degree in differential lncRNA-mRNA co-expression network.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Degree</th>
<th>Gene name</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRINITY_DN61993_c9_g1</td>
<td>29</td>
<td>Lnc-OM9822</td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN89620_c0_g1</td>
<td>25</td>
<td>Lnc-OM9785</td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN61638_c0_g1</td>
<td>18</td>
<td>Lnc-OM9852</td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN62639_c0_g4</td>
<td>16</td>
<td>Lnc-OM9744</td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN60962_c3_g3</td>
<td>16</td>
<td>Lnc-OM9748</td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN57241_c1_g3</td>
<td>15</td>
<td>Lnc-OM9787</td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN38889_c0_g1</td>
<td>14</td>
<td>Lnc-OM9826</td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN33630_c0_g1</td>
<td>14</td>
<td>Lnc-OM9694</td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN57241_c2_g3</td>
<td>14</td>
<td>ADP/ATP translocase 2</td>
<td>ADT2</td>
</tr>
<tr>
<td>TRINITY_DN60357_c0_g5</td>
<td>13</td>
<td>Lnc-OM9743</td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN58760_c7_g3</td>
<td>13</td>
<td>Lnc-OM9759</td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN46607_c0_g1</td>
<td>13</td>
<td>Lnc-OM9725</td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN48565_c0_g1</td>
<td>13</td>
<td>Sestrin 2</td>
<td>SESN2</td>
</tr>
<tr>
<td>TRINITY_DN63539_c4_g1</td>
<td>12</td>
<td>Actin alpha 1</td>
<td>ACTA1</td>
</tr>
<tr>
<td>TRINITY_DN59565_c3_g3</td>
<td>12</td>
<td>Calsequestrin-2</td>
<td>CASQ2</td>
</tr>
<tr>
<td>TRINITY_DN62931_c4_g5</td>
<td>12</td>
<td>Lnc-OM9693</td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN32746_c0_g2</td>
<td>12</td>
<td>Lnc-OM9780</td>
<td></td>
</tr>
<tr>
<td>TRINITY_DN62730_c1_g2</td>
<td>11</td>
<td>E3 ubiquitin-protein ligase</td>
<td>TRIM39</td>
</tr>
<tr>
<td>TRINITY_DN55042_c2_g2</td>
<td>11</td>
<td>Anion exchange protein 3</td>
<td>SLC4A3</td>
</tr>
<tr>
<td>TRINITY_DN60070_c2_g5</td>
<td>11</td>
<td>Lnc-OM9778</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1. The number of differentially expressed genes in rainbow trout under effect of E2. (A) Venn diagram of common differentially expressed mRNA in E2 treated fish between 2 time points (24hours post-injection VS 72hours post-injection). (B) Venn diagram of common differentially expressed lncRNA in E2 treated fish between 2 time points. (C) A volcano plot of differentially expressed transcripts (lncRNAs and mRNAs) between control and E2 treated group. (D-E) Unsupervised hierarchical clustering of the expression profiles of differentially expressed mRNAs (D) and lncRNAs (E) both distinguish E2 treated and control group.
Figure 2. Differentially expressed lncRNAs and mRNAs validated by qRT-PCR.

Comparison between RNA-seq and qRT-PCR validation results. X-axis shows lncRNAs (A) and mRNAs (B) validated in this study; Y-axis shows log$_2$ Ratio of expression of E2 versus control. Stars indicate significant difference based on t-test (p<0.05).
Figure 3. GO analysis of differentially expressed mRNAs.
(A) Treemap of up-regulated representative GO terms in BP. (B) Treemap of down-regulated representative GO terms in BP. (C) Treemap of up-regulated representative GO terms in MF. (D) Treemap of down-regulated representative GO terms in MF. (E) Treemap of up-regulated representative GO terms in CC. (F) Treemap of down-regulated representative GO terms in CC. Each rectangle is a single cluster representative. The representative GO terms were joined into ‘superclusters’ of loosely related terms with same color. BP indicates biological process. MF represents molecular function. CC indicates cellular component.
Figure 4. KEGG pathway analysis of differentially expressed genes. The rich factor was calculated using the gene count divided by the expect of gene count.
**Figure 5. IncRNA-mRNA co-expression network.** The differentially expressed IncRNA-mRNA co-expression network consisted of 681 co-expression relationships between 164 IncRNAs and 201 mRNAs. Blue circles represent IncRNAs, and red circles denote mRNAs.
Figure 6. IncRNA-pathway network. Blue circles represent IncRNAs, and orange circles denote pathways.
Figure 7. Key lncRNAs and corresponding crucial pathways. (A) Sub-network of key lncRNAs and their regulatory pathways. (B) Estrogen signaling pathway. Red rectangles indicate proteins, which are differentially expressed and potentially regulated by lncRNAs. (C) Estrogen response element (ERE) located in lncRNA (Lnc-OM9822). (D) The interaction matrix that shows protein-RNA interaction between estrogen receptor alpha and Lnc_OM9822. The shades of red of the heat-map indicate the interaction score of individual amino acid and nucleotide pairs.
Figure 8. Putative motifs across sequence of lnc-OM9822. (A) The sequence of five motifs found in lnc-OM9822. (B) The position of all motifs on lnc-OM9822.
Identification of long intergenic non-coding RNAs (lincRNAs) in bovine oocyte using next-generation sequencing

Jian Wang\textsuperscript{1}, Prasanthi P. Koganti\textsuperscript{1}, and Jianbo Yao\textsuperscript{1}\textsuperscript{*}

\textsuperscript{1}Division of Animal and Nutritional Sciences, West Virginia University, Morgantown, WV 26506 USA

\textsuperscript{*}Corresponding author
Accumulating evidence suggests that lncRNAs are key regulators of diverse cellular processes. Recent advances in RNA sequencing (RNA-Seq) and bioinformatics methods have provided an opportunity to identify and analyze such transcripts. Maternal transcripts that accumulate in the oocyte during oogenesis play important roles during initial stages of embryonic development. LncRNAs have been reported to regulate embryonic development in human and mouse. However, very little is known about the identity and characteristics of lncRNAs in bovine oocytes. In the present study, we performed ab initio assembly of more than 80 million RNA-Seq reads from oocytes, and identified 1,535 transcribed lncRNAs from 1,183 loci. In addition, comparing with previous studies and NONCODE database, we found 115 (7.6%) of our lincRNAs overlap with previously reported bovine lncRNAs in NONCODE database and 565 reported in previous papers. Furthermore, we calculated tissue specificity score for each oocyte lincRNA. The results indicate that the majority of oocyte lincRNAs (80%) are tissue-specific. Finally, functional prediction of oocyte-specific lincRNAs suggested their involvement in oogenesis through regulating their neighboring protein-coding genes. This study provides a starting point for future research aimed at understanding the roles of lncRNAs in controlling oocyte development and early embryogenesis in cattle.
INTRODUCTION

Over the past decade, genome-wide transcriptional studies discovered that the vast majority of the mammalian genome (up to 80%) is transcribed, while only 2-3% of the mammalian genome is transcribed into protein-coding RNAs (mRNAs). The majority of transcripts function as non-translated RNA molecules called non-coding RNA (ncRNA). Recent advances in transcriptome sequencing have allowed for the discovery of a new class of ncRNA that are generally longer than 200 nucleotides, known as long non-coding RNA (lncRNA). LncRNA transcribed from intergenic region is referred to as lincRNA. Like protein-coding genes, lncRNAs are usually 5’ capped, 3’ polyadenylated and alternative spliced. The study of lncRNA is now focusing on understanding their functions, revealing that lncRNA play various roles in diverse biological processes, including regulation of epigenetic marks and gene expression on different levels, as well as protein post translational modification. According to the genomic position of the loci from which they are transcribed and their proximity to protein coding genes in the genome, lncRNAs can be divided into five categories: sense, antisense, bidirectional, intronic and intergenic lncRNAs. In the past few years, an increasing number of lncRNAs have been reported in eukaryotic organisms, ranging from nematodes to human. So far, three major lncRNA database including LNCipedia, GENCODE and NONCODE have identified more than 100,000 human lncRNA genes.

Cattle (Bos taurus) is one of the most commonly raised livestock for meat, milk and other dairy products. There are a number of studies reporting bovine lncRNA across many tissues. Because of the key role of lncRNAs in regulation of gene expression, it is important to identify all lncRNAs in cattle. A total of 449 putative lncRNAs have been identified using public bovine expressed sequence tags sequences. Moreover, more than 4,000 lncRNAs were predicted in bovine skin using RNA-Seq data. Billery et al. (2014) identified a stringent set of 584 lincRNAs in bovine muscle. More recently, Koufariotis et al. (2015) reported a total of 9,778 lncRNAs from RNA-Seq data across 18 tissues. However, very little is known about the identity and characteristics of lncRNAs in bovine oocytes. The developmental competence of an oocyte, also known as egg quality, is defined as the ability of the egg to be fertilized and subsequently develop into a normal
embryo. Mammalian oocytes become transcriptionally silent following germinal-vesicle breakdown, so the final stages of oocyte maturation and early embryo development depend on stored transcripts. Several studies have reported that lncRNA plays critical roles in the embryonic stem cell regulatory\textsuperscript{6,18-20}. For instance, more than 100 lncRNA promoters were identified to be bound by stem cell factors such as OCT4 and Nanog\textsuperscript{20}. Therefore, the study of lncRNA in bovine oocyte could help us to understand the embryonic development. In this study we described a comprehensive catalogue of putative lincRNA expressed in bovine oocytes. We also compared our results to those from other bovine studies, assessed the tissue specificity of each lincRNA and performed function prediction for oocyte specific linRNAs.
MATERIALS AND METHODS

RNA-seq data

RNA-seq data for transcriptome reconstruction in this study includes polyadenylated RNA sample from bovine GV (germinal vesicle) and MII (metaphase II) stage oocytes sequenced by the Illumina Genome Analyzer II (GAII). A total of 85 million reads were yielded, and the reads were paired and both lengths were 100 bp.

Publicly available annotations

Protein-coding genes were downloaded from UCSC genome browser\textsuperscript{21} and Ensembl genome browser\textsuperscript{22}. All known noncoding genes were downloaded from Ensembl genome browser\textsuperscript{22} and NONCODE database\textsuperscript{12}.

RNA-seq reads mapping and assembly

After trimming adaptor sequences and filtering rRNA and ambiguous and low quality bases, a total of 78 million pair-end reads were obtained. Spliced read aligner TopHat2 was used to align all clean reads to the bovine genome (UMD3.1) using the default parameters. Aligned reads from TopHat2 were assembled into transcriptome by Scripture\textsuperscript{6} and Cufflinks\textsuperscript{23}. Both assemblers use spliced read information to determine exons connectivity; however, with two different approaches. Cuffcompare\textsuperscript{23} was used to determine a unique set of isoforms assembled from both assemblers for further lincRNAs identification.

LincRNA identification pipeline

Identification of each transcript as either coding or noncoding was performed using a step-wise pipeline to filter out the transcripts that had a high chance of being protein coding.

First, we eliminated all transcripts that had exon overlapping a transcript from any of the following sets: (1) coding genes annotated in UCSC, RefSeq and Ensembl, (2) microRNA, tRNAs, snoRNAs, rRNAs annotated in Ensembl. Finally, we selected
unknown transcripts with class code of “u”.

Second, coding potential of each candidate transcript were assessed using PhyloCSF\textsuperscript{24} and CPAT\textsuperscript{25}. PhyloCSF uses a multispecies nucleotide sequence alignment to estimate the degree of evolutionary pressure on sequence substitutions to preserve an open reading frame. We ran PhyloCSF using multiple sequence alignment of 5 mammalian genomes including cow, human (hg19), mouse (mm10), rat (m5) and dog (canfam3). In addition, CPAT is another program to assess the coding potential, which is based on the length and quality of ORF (open reading frame) to assess a transcript’s coding potential with premise that true protein coding gene is more likely to have a long and high-quality ORF.

Third, to evaluate which of the remaining transcripts contains a known protein coding domain, HMMER-3\textsuperscript{26} was used to identify transcripts translated in all three possible frames having homologs with any of the 31,912 known protein family domains in the Pfam database (release 24; both PfamA and PfamB). All transcripts with a Pfam hit were excluded.

Finally, putative protein-coding RNAs were filtered out by applying a maximal ORF (Open reading frame) length threshold. All transcripts with a maximal ORF > 100 amino acids were excluded.

Tissue specificity score

To evaluate tissue specificity of a transcript, an entropy-based metric that relies on Jensen-Shannon (JS) divergence was used to calculate specificity scores (0 to 1). A perfect tissue-specific pattern is scored as JS=1, which means a transcript is expressed only in one tissue\textsuperscript{27}. 
RESULTS

Transcriptome reconstruction of bovine oocytes

To prepare for sequencing, RNAs were extracted from bovine oocytes. A total of 85 million raw reads were generated from sequencing. These reads were paired and both lengths were 100 bp. After quality control, 78 million clean reads were obtained (Table 1). All clean reads were further mapped to bovine genome (UMD3.1) using TopHat2\(^{28}\). 78.4% (61 million) of the clean reads were aligned onto the bovine genome, and 82% of the mapped reads were aligned concordantly. The mapping ratio was similar to those obtained in other RNA-Seq bovine studies\(^{29-34}\). We then used *ab initio* assembly software Scripture\(^6\) and Cufflinks\(^{23}\) to reconstruct the transcriptome based on the read-mapping results. Transcripts reconstructed by these two assemblers were merged into a combined set of transcripts using the Cuffcompare utility provided by Cufflinks, resulting in the assembly of a total number of 42,396 transcripts from 37,678 genomic loci.

All assembled transcripts were categorized using the bovine genome annotation obtained from UCSC and Ensembl genome browser (Table 1). Approximately 40% of the transcripts correspond to already annotated transcripts. Notably, ~17% (7106) of the transcripts correspond to novel isoforms of known genes (“j” class), indicating that a large number of new transcript isoforms have yet to be annotated or the bovine genome remains poorly annotated. Interestingly, more than 19% (8336) of the transcripts were categorized as unknown intergenic transcripts (“u” class). After removing all single exon unknown intergenic transcripts, 2552 multi exon transcripts that have class code of “u” were selected for lincRNA identification.

Identification of putative lincRNAs

To identify lincRNAs, we first analyzed the coding potential of all 2,552 novel intergenic transcripts using CPAT\(^{25}\) and PhyloCSF\(^{24}\). PhyloCSF scores were first calculated for the 2,552 putative multi exon intergenic transcripts. All transcripts with a negative score were retained as potential non-coding candidates. In addition, CPAT was also used to assess the coding potential for all 2,552 transcripts. To determine the
optimum cut-off value, CPAT was trained using a set of 10,000 bovine CDS from Refseq, a set of 3,650 ncRNAs from Ensembl and a set of 6,350 intron sequences from Refseq. A cut-off value of 0.348 was selected. Any transcript predicted by both CPAT and PhyloCSF as non-coding RNA was selected as potential bovine lincRNA. This procedure identified 1,627 transcripts from 1249 different genome loci. Finally, we scanned each remaining transcripts and classified whether it contained any of the known protein coding domains in the Pfam database. This step filtered out 92 transcripts and resulted in 1,535 bovine lincRNAs corresponding to 1,183 putative non-coding genes.

Previous studies in mammals have shown that lncRNAs are shorter, and have fewer exon number compared with protein-coding genes. In the present study, the mean length and average exon number of bovine oocyte lincRNA are 782 +/- 580 nt and 2.6 +/- 0.8 exons, which is similar with human (~1000 nt and 2.9 exons) and zebrafish (~1000 nt and 2.8 exons). The genome distribution of putative lincRNAs was also investigated. As shown in Fig. 1, chromosome 7 has the grease number of bovine oocyte lincRNA, followed by chromosome 10, 1, X, 8 and 2, whereas, chromosome 15 has least number of lincRNA.

Comparative analysis with bovine IncRNAs from similar studies

Comparison of the genomic position of the 1,535 bovine lincRNAs with the position of previously identified IncRNAs in the NONCODEv4 database show that 115 (7.6%) of our lincRNAs overlap with previously reported bovine IncRNAs.

A comparison of lincRNA identified in this study with the IncRNA found in similar studied was also performed (Fig. 2). We obtained the IncRNA list from the studies by LT. Koufariotis et al.¹⁷, C. Billerey et al.¹⁶ and R. Weikard et al.¹⁵. Weikard et al.¹⁵ identified 4,899 IncRNA, of which 63 IncRNAs were found to overlap with oocyte lincRNAs. Moreover, a total of 55 (out of 584) IncRNAs were found to overlap with our putative oocyte lincRNAs. Furthermore, a total of 9,778 transcripts were identified as IncRNA across 18 bovine tissues, including 506 IncRNAs also found in our study. Notably, only 2 IncRNAs were shared by all four studies. 970 lincRNAs were unique to bovine oocyte. All these results indicate that IncRNAs were expressed in a tissue-specific manner.
Tissue specificity of bovine oocyte lincRNAs

In order to calculate the tissue specificity score for each oocyte lincRNA, we downloaded RNA-Seq data sets of other 9 bovine tissues from NCBI SRA database (Accession number SRR594491- SRR594499). The FPKM (fragments per kilobase of transcript per million mapped reads) value of each transcript in each of 9 tissues was calculated by Cufflinks. We then calculated a tissue specificity score for each lincRNA transcript using an entropy-based metric that relies on the Jensen-Shannon (JS) divergence\textsuperscript{27}. The distribution of JS scores was shown in Fig. 3. Using a JS score of 0.5 as a cutoff\textsuperscript{36}, the majority of oocyte lincRNAs (80%) are tissue-specific. Notably, more than 37\% of oocyte lincRNAs had a JS score of 1, demonstrating they are expressed exclusively in bovine oocyte.

Functional prediction of bovine oocyte lincRNAs

Recent studies suggest that some lincRNAs may act in \textit{cis} and regulate the expression of a neighboring protein coding gene\textsuperscript{37,38}. The expectation of the \textit{cis} regulation hypothesis is that the expression between lincRNA and its neighboring gene would be correlated across all samples used in the present study. In this study, we first selected lincRNAs showing oocyte specific expression (JS score of 0.5). A total of 1239 lincRNAs have a JS score larger than 0.5. Moreover, we then screened \textasciitilde50 Kb genomic region as neighboring chromosome region\textsuperscript{39} flanking the genomic loci of all 1239 lincRNAs on either direction using BEDTools\textsuperscript{40}. Finally, 202 mRNAs were identified as “neighbor” of bovine oocyte lincRNAs. Furthermore, we calculated the Pearson’s correlation coefficients (PCC) between lincRNAs and their neighboring genes and analyzed enriched GO terms associated with mRNAs that are strongly correlated with neighboring lincRNA. Finally, we identified 75 oocyte specific lincRNAs strongly co-expressed with 58 neighboring protein-coding genes. As the results shown in Fig. 4 and Supplemental Fig. S1, mitochondrial respiratory chain complex assembly, cytoskeleton organization, protein modification, microtubule-based process were enriched in biological process (Fig. 4A). Zinc ion transmembrane transporter activity, phosphatase regulator activity and nucleosomal DNA binding were over-represented in molecular function (Fig. 4B). The
enrichment of neighboring genes in cellular component most related to phosphatase complex, condensed chromosome outer kinetochore, CCAAT-binding factor complex, pericentriolar material and spindle microtubule (Fig. 4C). These results suggest that a portion of bovine oocyte lincRNAs might act locally to regulate their neighboring genes in cis.
DISCUSSION

In this study, we presented the comprehensive annotation of bovine oocyte lincRNAs using polyA(+) captured RNA-Seq data from bovine GV and MII stage oocytes. We first reconstructed bovine oocyte transcriptome from deep sequencing data to reveal a significant number of novel lincRNAs. In order to assess the tissue specificity of newly identified lincRNAs in bovine oocyte, we collected RNA-Seq data sets from multiple bovine tissues from NCBI RSA database. The tissue-specificity score was calculated based on the FPKM for each transcript and demonstrated that bovine oocyte lincRNAs are expressed in a much more tissue-specific manner. Furthermore, based on the hypothesis that lincRNAs might act in cis to regulate the gene expression in their chromosomal neighborhood, we were able to predict the putative functions for 75 lincRNAs.

Most importantly, we found that cytoskeleton organization, regulation of microtubule-based process, zinc ion transport and mitochondrial respiratory chain complex assembly were over-represented for neighboring genes of oocyte specific lincRNAs. Early embryonic development in many organisms relies on the subcellular organization of the oocyte and requires the coordination of a variety of cellular events. Cytoskeleton was believed to mediate many of these processes. More importantly, microtubules, a component of the cytoskeleton, are the major constituents of spindles that are used to pull apart eukaryotic chromosomes during mitosis and meiosis. Moreover, Stephenson et al.\textsuperscript{41} reported that the zinc level would affect bovine oocyte maturation and fertilization in vitro. Furthermore, Kong et al. demonstrate that rapid cellular zinc influx regulates early mammalian development during the oocyte-to-egg transition through modulation of the meiotic cell cycle\textsuperscript{42}. Instead of transcriptionally based mechanism, they found that it is two maternally derived zinc transporter that control zinc uptake. Targeted knockdown of these transporters during meiotic maturation perturbs the intracellular zinc quota and results in a cell cycle arrest at a telophase I-like state in mouse oocyte. The importance of mitochondria was highlighted by their crucial role to support critical events such as spindle formation, chromatid separation and cell division during oocyte maturation. It is known that the developing zygote is dependent on the existing pool of
mitochondria until blastocyst implantation\textsuperscript{43}.

Reversible phosphorylation is important in regulating oocyte meiosis. The inhibition of phosphatase-1 (PP1) and PP2A was found to stimulate oocyte germinal vesicle breakdown\textsuperscript{44}. Phosphorylation of PP1 at Thr320 by cyclin dependent kinase-1 (CDK1) causes PP1 inactivation. Germinal vesicle-intact oocytes did not contain phosphorylation of Thr320 of PP1. Moreover, inhibition of oocyte germinal vesicle breakdown by roscovitine (ROSC) was shown to coincide with PP1 phosphorylation at Thr320\textsuperscript{45}. Besides, the pericentriolar material (PCM) is a matrix of proteins serving as a platform for spindle assembly\textsuperscript{46}. The over representation of PCM, together with the enrichment of condensed chromosome outer kinetochore and spindle microtubule, suggest activity of spindle apparatus assembly. Taken together, these results indicate the involvement of bovine oocyte lincRNAs in oogenesis through regulating their neighboring protein-coding genes.
CONCLUSION

In the present study, we performed an initial assembly of more than 80 million RNA-Seq reads from bovine GV and MII stage oocytes and identified 1,535 transcribed lincRNAs from 1,183 loci. In addition, comparing with previous studies and NONCODE database, we found 115 (7.6%) of our lincRNAs overlap with previously reported bovine lncRNAs in NONCODE database and 565 reported in previous papers. Furthermore, we calculated the tissue specificity score for each oocyte lincRNA. The results indicate that the majority of oocyte lincRNAs (80%) are tissue-specific. Finally, we propose function of oocyte specific lincRNAs, which suggested the involvement of bovine oocyte lincRNAs in oogenesis through regulating their neighboring protein-coding genes. This study provides a starting point for future studies aimed at understanding the roles of lncRNAs in controlling oocyte development and early embryogenesis in cattle.
REFERENCES


Wickramasinghe, S., Rincon, G., Islas-Trejo, A. & Medrano, J. F. Transcriptional profiling of bovine milk using RNA sequencing. *BMC genomics* 13, 45,


Kong, B. Y. et al. Maternally-derived zinc transporters ZIP6 and ZIP10 drive the


Table 1. Statistical summary of bovine oocyte sequencing, assembly and annotation

<table>
<thead>
<tr>
<th>RNA-sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of raw reads</td>
</tr>
<tr>
<td>Number of clean reads</td>
</tr>
<tr>
<td>Number of mapped reads</td>
</tr>
<tr>
<td>Number of concordant pair alignment</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of transcripts in each transfrag class</th>
</tr>
</thead>
<tbody>
<tr>
<td>=</td>
</tr>
<tr>
<td>c</td>
</tr>
<tr>
<td>j</td>
</tr>
<tr>
<td>e</td>
</tr>
<tr>
<td>i</td>
</tr>
<tr>
<td>o</td>
</tr>
<tr>
<td>p</td>
</tr>
<tr>
<td>u</td>
</tr>
<tr>
<td>x</td>
</tr>
<tr>
<td>s</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LincRNA identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of novel transcripts with multiple exon</td>
</tr>
<tr>
<td>Number of transcripts without coding potential</td>
</tr>
<tr>
<td>Number of lincRNA (protein domain filter)</td>
</tr>
<tr>
<td>Average length</td>
</tr>
<tr>
<td>Number of average exon</td>
</tr>
</tbody>
</table>

Figure 1. Total number of putative oocyte lincRNA per chromosome.
Figure 2. Venn diagram of comparative analysis of lincRNA in present study with bovine lncRNA from similar studies. The green circle represents the lncRNA found in bovine muscle. The blue circle represents the lncRNA found in bovine skin. The orange circle represents the lncRNA identified in 18 bovine tissues not including oocyte.
Figure 3. Tissue-specific expression of oocyte lincRNAs.
Figure 4. The GO analysis of neighboring mRNAs. (A) Treemap of representative GO terms in BP. (B) Treemap of representative GO terms in MF. (C) Treemap of representative GO terms in CC.
Summary
Summary

The non-coding RNA that has attracted the attentions of many researchers recently is long non-coding RNA (lncRNA). The whole-transcriptome analysis revealed vast number of lncRNAs that involved in various biological functions. Yet most of them are not annotated and the functional and molecular mechanisms have not been known. As the numbers of discovered lncRNAs are huge, it is difficult to test the functionality of lncRNA in the laboratory for all of them. Hence, the computational method is mainly used to predict the functionality of different lncRNAs. The goal of this study was to identify lncRNA in rainbow trout and cattle and to predict their functions.

The first study describe the first reference catalog of 9,674 rainbow trout lincRNAs based on analysis of RNA-Seq data from 15 tissues. Systematic analysis revealed that lincRNAs in rainbow trout share many characteristics with those in other mammalian species. Co-expression network analysis suggested that many lincRNAs are associated with immune response, muscle differentiation and neural development. This study provides an opportunity for future experimental and computational studies to uncover the functions of lincRNAs in rainbow trout.

In the second chapter, we performed next-generation RNA sequencing and comprehensive bioinformatics analyses to characterize the transcriptome profiles, including mRNAs and lncRNAs, in skeletal muscle of normal and E2 treated rainbow trout. A total of 226 lncRNAs and 253 mRNAs were identified as differentially regulated. We identified crucial pathways, including several signal transduction pathways, hormone response, oxidative response and protein, carbon and fatty acid metabolism pathways. Subsequently, a functional lncRNA-mRNA co-expression network was constructed, which consisted of 681 co-expression relationships between 164 lncRNAs and 201 mRNAs. Moreover, an lncRNA-pathway network was constructed. A total of 65 key lncRNAs were identified, which regulate 20 significantly enriched pathways. Finally, the function of a novel lncRNA (lnc-OM9822) was predicted, which may be activated by estrogen receptor alpha (ER1) and involve in the estrogen-signaling pathway. Overall, our analysis not only effectively provides insights into the mRNA and lncRNA association with the effect of E2 on skeletal muscle in rainbow trout but also provides further insights into understanding the molecular mechanism of lncRNAs.
In the last chapter, we performed ab initio assembly of more than 80 million RNA-Seq reads from oocytes, and identified 1,535 transcribed lncRNAs from 1,183 loci. In addition, comparing with previous studies and NONCODE database, we found 115 (7.6%) of our lincRNAs overlap with previously reported bovine lincRNAs in NONCODE database and 565 reported in previous papers. Furthermore, we calculated tissue specificity score for each oocyte lincRNA. The results indicate that the majority of oocyte lincRNAs (80%) are tissue-specific. Finally, we predicted function of oocyte specific lincRNAs suggested the involvement of bovine oocyte lincRNAs in oogenesis through regulating their neighboring protein-coding genes. This study provides a starting point for future research aimed at understanding the roles of lncRNAs in controlling oocyte development and early embryogenesis in cattle.
REFERENCES


APPENDIX

CHAPTER 1 - Supplemental files


Supplemental file 6: Genes and their memberships to each module. MM stands for module membership, which is the correlation between a gene and a module. PMM is the p-value of MM. (https://static-content.springer.com/esm/art%3A10.1007%2Fs10126-016-9689-5/MediaObjects/10126_2016_9689_MOESM4_ESM.xlsx)

Supplemental file 7: GO analysis of modules. GO terms enriched in specific modules were tested by Fisher’s exact test. (https://static-content.springer.com/esm/art%3A10.1007%2Fs10126-016-9689-5/MediaObjects/10126_2016_9689_MOESM5_ESM.xlsx)
CHAPTER 2 - Supplemental files

Supplemental Figure S1. Plots of sample relations of time point of 24 Hours. Multidimensional scaling plot show the relationship between all pairs of samples.
Supplemental Figure S2. Directed acyclic graph (DAG) of over-represented and hormone response associated GO terms.
Supplemental Table S1. Key pathways in E2 treated skeletal muscle.

<table>
<thead>
<tr>
<th>Regulatory IncRNAs (n)</th>
<th>Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>Focal adhesion</td>
</tr>
<tr>
<td>28</td>
<td>Regulation of actin cytoskeleton</td>
</tr>
<tr>
<td>17</td>
<td>Cardiac muscle contraction</td>
</tr>
<tr>
<td>15</td>
<td>Calcium signaling pathway</td>
</tr>
<tr>
<td>15</td>
<td>Rap1 signaling pathway</td>
</tr>
<tr>
<td>13</td>
<td>PPAR signaling pathway</td>
</tr>
<tr>
<td>10</td>
<td>Adipocytokine signaling pathway</td>
</tr>
<tr>
<td>9</td>
<td>AMPK signaling pathway</td>
</tr>
<tr>
<td>7</td>
<td>Carbon metabolism</td>
</tr>
<tr>
<td>6</td>
<td>Jak-STAT signaling pathway</td>
</tr>
<tr>
<td>6</td>
<td>Protein processing in endoplasmic reticulum</td>
</tr>
<tr>
<td>5</td>
<td>Arginine and proline metabolism</td>
</tr>
<tr>
<td>5</td>
<td>Fatty acid metabolism/degradation</td>
</tr>
<tr>
<td>5</td>
<td>Beta-Alanine metabolism</td>
</tr>
<tr>
<td>4</td>
<td>Citrate cycle (TCA cycle)</td>
</tr>
<tr>
<td>4</td>
<td>Protein digestion and absorption</td>
</tr>
<tr>
<td>3</td>
<td>Regulation of autophagy</td>
</tr>
<tr>
<td>2</td>
<td>Estrogen signaling pathway</td>
</tr>
<tr>
<td>2</td>
<td>Insulin resistance</td>
</tr>
<tr>
<td>2</td>
<td>Pyruvate metabolism</td>
</tr>
</tbody>
</table>
## Supplemental Table S2. Key lncRNAs in E2 treated skeletal muscle.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Gene ID</th>
<th>Regulatory pathways (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lnc-OM9822</td>
<td>TRINITY_DN61993_c9_g1</td>
<td>8</td>
</tr>
<tr>
<td>Lnc-OM9826</td>
<td>TRINITY_DN38889_c0_g1</td>
<td>7</td>
</tr>
<tr>
<td>Lnc-OM9787</td>
<td>TRINITY_DN57241_c1_g3</td>
<td>7</td>
</tr>
<tr>
<td>Lnc-OM9780</td>
<td>TRINITY_DN32746_c0_g2</td>
<td>6</td>
</tr>
<tr>
<td>Lnc-OM9731</td>
<td>TRINITY_DN45834_c0_g1</td>
<td>6</td>
</tr>
<tr>
<td>Lnc-OM9832</td>
<td>TRINITY_DN59109_c0_g4</td>
<td>6</td>
</tr>
<tr>
<td>Lnc-OM9887</td>
<td>TRINITY_DN59603_c2_g1</td>
<td>6</td>
</tr>
<tr>
<td>Lnc-OM9814</td>
<td>TRINITY_DN61890_c2_g1</td>
<td>6</td>
</tr>
<tr>
<td>Lnc-OM9714</td>
<td>TRINITY_DN61931_c5_g1</td>
<td>6</td>
</tr>
<tr>
<td>Lnc-OM9786</td>
<td>TRINITY_DN64001_c9_g4</td>
<td>6</td>
</tr>
<tr>
<td>Lnc-OM9785</td>
<td>TRINITY_DN89620_c0_g1</td>
<td>6</td>
</tr>
<tr>
<td>Lnc-OM9727</td>
<td>TRINITY_DN49942_c1_g1</td>
<td>5</td>
</tr>
<tr>
<td>Lnc-OM9809</td>
<td>TRINITY_DN54235_c1_g1</td>
<td>5</td>
</tr>
<tr>
<td>Lnc-OM9761</td>
<td>TRINITY_DN58884_c6_g1</td>
<td>5</td>
</tr>
<tr>
<td>Lnc-OM9693</td>
<td>TRINITY_DN62931_c4_g5</td>
<td>5</td>
</tr>
<tr>
<td>Lnc-OM9720</td>
<td>TRINITY_DN51283_c1_g1</td>
<td>4</td>
</tr>
<tr>
<td>Lnc-OM9733</td>
<td>TRINITY_DN61361_c3_g1</td>
<td>4</td>
</tr>
<tr>
<td>Lnc-OM9734</td>
<td>TRINITY_DN61850_c1_g1</td>
<td>4</td>
</tr>
<tr>
<td>Lnc-OM9735</td>
<td>TRINITY_DN62241_c2_g6</td>
<td>4</td>
</tr>
<tr>
<td>Lnc-OM9724</td>
<td>TRINITY_DN50458_c2_g1</td>
<td>3</td>
</tr>
<tr>
<td>Lnc-OM9820</td>
<td>TRINITY_DN58167_c3_g1</td>
<td>3</td>
</tr>
<tr>
<td>Lnc-OM9695</td>
<td>TRINITY_DN58349_c1_g3</td>
<td>3</td>
</tr>
<tr>
<td>Lnc-OM9759</td>
<td>TRINITY_DN58760_c7_g3</td>
<td>3</td>
</tr>
<tr>
<td>Lnc-OM9768</td>
<td>TRINITY_DN60390_c2_g6</td>
<td>3</td>
</tr>
<tr>
<td>Lnc-OM9781</td>
<td>TRINITY_DN60546_c2_g3</td>
<td>3</td>
</tr>
<tr>
<td>Lnc-OM9690</td>
<td>TRINITY_DN60550_c4_g2</td>
<td>3</td>
</tr>
<tr>
<td>Lnc-OM9748</td>
<td>TRINITY_DN60962_c3_g3</td>
<td>3</td>
</tr>
<tr>
<td>Lnc-OM9805</td>
<td>TRINITY_DN61439_c1_g1</td>
<td>3</td>
</tr>
<tr>
<td>Lnc-OM9871</td>
<td>TRINITY_DN61485_c2_g2</td>
<td>3</td>
</tr>
<tr>
<td>Lnc-OM9710</td>
<td>TRINITY_DN61788_c0_g1</td>
<td>3</td>
</tr>
<tr>
<td>Lnc-OM9716</td>
<td>TRINITY_DN62274_c3_g3</td>
<td>3</td>
</tr>
<tr>
<td>Lnc-OM9744</td>
<td>TRINITY_DN62639_c0_g4</td>
<td>3</td>
</tr>
<tr>
<td>Lnc-OM9708</td>
<td>TRINITY_DN118820_c1_g2</td>
<td>2</td>
</tr>
<tr>
<td>Lnc-OM9679</td>
<td>TRINITY_DN13989_c0_g1</td>
<td>2</td>
</tr>
<tr>
<td>Lnc-OM9725</td>
<td>TRINITY_DN46607_c0_g1</td>
<td>2</td>
</tr>
<tr>
<td>Lnc-OM9784</td>
<td>TRINITY_DN54282_c1_g2</td>
<td>2</td>
</tr>
<tr>
<td>Lnc-OM9881</td>
<td>TRINITY_DN55879_c11_g11</td>
<td>2</td>
</tr>
<tr>
<td>Lnc-OM9732</td>
<td>TRINITY_DN57669_c4_g1</td>
<td>2</td>
</tr>
<tr>
<td>Lnc-OM9687</td>
<td>TRINITY_DN58453_c0_g1</td>
<td>2</td>
</tr>
<tr>
<td>Lnc-OM9742</td>
<td>TRINITY_DN58760_c7_g1</td>
<td>2</td>
</tr>
<tr>
<td>Lnc-OM9852</td>
<td>TRINITY_DN61638_c0_g1</td>
<td>2</td>
</tr>
<tr>
<td>Lnc-OM9711</td>
<td>TRINITY_DN62878_c4_g8</td>
<td>2</td>
</tr>
<tr>
<td>Lnc-OM9765</td>
<td>TRINITY_DN30100_c0_g1</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9694</td>
<td>TRINITY_DN33630_c0_g1</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9753</td>
<td>TRINITY_DN50001_c1_g2</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9861</td>
<td>TRINITY_DN54010_c0_g1</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9804</td>
<td>TRINITY_DN56675_c3_g1</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9704</td>
<td>TRINITY_DN57623_c6_g3</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9789</td>
<td>TRINITY_DN57732_c0_g2</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9751</td>
<td>TRINITY_DN58681_c3_g1</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9823</td>
<td>TRINITY_DN58681_c3_g5</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9706</td>
<td>TRINITY_DN59071_c1_g1</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9775</td>
<td>TRINITY_DN59937_c11_g2</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9778</td>
<td>TRINITY_DN60070_c2_g5</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9722</td>
<td>TRINITY_DN60750_c8_g1</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9691</td>
<td>TRINITY_DN61246_c7_g1</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9894</td>
<td>TRINITY_DN61494_c2_g1</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9847</td>
<td>TRINITY_DN62401_c7_g1</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9723</td>
<td>TRINITY_DN62730_c0_g2</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9856</td>
<td>TRINITY_DN62897_c1_g5</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9892</td>
<td>TRINITY_DN63148_c7_g9</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9684</td>
<td>TRINITY_DN63884_c1_g2</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9818</td>
<td>TRINITY_DN64083_c2_g1</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9797</td>
<td>TRINITY_DN64112_c0_g2</td>
<td>1</td>
</tr>
<tr>
<td>Lnc-OM9678</td>
<td>TRINITY_DN90140_c0_g1</td>
<td>1</td>
</tr>
</tbody>
</table>
Supplemental Figure 1. The GO analysis of neighboring mRNAs. (A) Scatterplot of representative GO terms in BP. (B) Scatterplot of representative GO terms in MF. (C) Scatterplot of representative GO terms in CC